

Inference for High-dimensional Differential Correlation Matrices ^{*}

T. Tony Cai and Anru Zhang[†]

Department of Statistics

The Wharton School

University of Pennsylvania

Abstract

Motivated by differential co-expression analysis in genomics, we consider in this paper estimation and testing of high-dimensional differential correlation matrices. An adaptive thresholding procedure is introduced and theoretical guarantees are given. Minimax rate of convergence is established and the proposed estimator is shown to be adaptively rate-optimal over collections of paired correlation matrices with approximately sparse differences. Simulation results show that the procedure significantly outperforms two other natural methods that are based on separate estimation of the individual correlation matrices. The procedure is also illustrated through an analysis of a breast cancer dataset, which provides evidence at the gene co-expression level that several genes, of which a subset has been previously verified, are associated with the breast cancer. Hypothesis testing on the differential correlation matrices is also considered. A test, which is particularly well suited for testing against sparse alternatives, is introduced. In addition, other related problems, including estimation of a single sparse correlation matrix, estimation of the differential covariance matrices, and estimation of the differential cross-correlation matrices, are also discussed.

Keywords: Adaptive thresholding, covariance matrix, differential co-expression analysis, differential correlation matrix, optimal rate of convergence, sparse correlation matrix, thresholding.

^{*}The research was supported in part by NSF Grant DMS-1208982 and NIH Grant R01 CA127334.

[†]Corresponding author. E-mail: anrzhang@wharton.upenn.edu

1 Introduction

Statistical inference on the correlation structure has a wide array of applications, ranging from gene co-expression network analysis (Carter et al., 2004; Lee et al., 2004; Zhang et al., 2008; Dubois et al., 2010; Fuller et al., 2007) to brain intelligence analysis (Shaw et al., 2006). For example, understanding the correlations between the genes is critical for the construction of the gene co-expression network. See Kostka and Spang (2004), Lai et al. (2004), and Fuller et al. (2007). Driven by these and other applications in genomics, signal processing, empirical finance, and many other fields, making sound inference on the high-dimensional correlation structure is becoming a crucial problem.

In addition to the correlation structure of a single population, the difference between the correlation matrices of two populations is of significant interest. Differential gene expression analysis is widely used in genomics to identify disease-associated genes for complex diseases. Conventional methods mainly focus on the comparisons of the mean expression levels between the disease and control groups. In some cases, clinical disease characteristics such as survival or tumor stage do not have significant associations with gene expression, but there may be significant effects on gene co-expression related to the clinical outcome (Shedden and Taylor (2005); Hudson et al. (2009); Bandyopadhyay et al. (2010)). Recent studies have shown that changes in the correlation networks from different stages of disease or from case and control groups are also of importance in identifying dysfunctional gene expressions in disease. See, for example, de la Fuente (2010). This differential co-expression network analysis has become an important complement to the original differential expression analysis as differential correlations among the genes may reflect the rewiring of genetic networks between two different conditions (See Shedden and Taylor (2005); Bandyopadhyay et al. (2010); de la Fuente (2010); Ideker and Krogan (2012); Fukushima (2013)).

Motivated by these applications, we consider in this paper optimal estimation of the differential correlation matrix. Specifically, suppose we observe two independent sets of p -dimensional i.i.d. random samples $\mathbf{X}^{(t)} = \{\mathbf{X}_1^{(t)}, \dots, \mathbf{X}_{n_t}^{(t)}\}$ with mean $\boldsymbol{\mu}_t$, covariance matrix $\boldsymbol{\Sigma}_t$, and correlation matrix \mathbf{R}_t , where $t = 1$ and 2 . The goal is to estimate the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$. A particular focus of the paper is on estimating an approximately sparse differential correlation matrix in the high dimensional setting where the dimension is much larger

than the sample sizes, i.e., $p \gg \max(n_1, n_2)$. The estimation accuracy is evaluated under both the spectral norm loss and the Frobenius norm loss.

A naive approach to estimating the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ is to first estimate the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ separately and then normalize to obtain estimators $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$ of the individual correlation matrices \mathbf{R}_1 and \mathbf{R}_2 , and finally take the difference $\hat{\mathbf{D}} = \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ as the estimator of the differential correlation matrix \mathbf{D} . A simple estimate of a correlation matrix is the sample correlation matrix. However, in the high-dimensional setting, the sample correlation matrix is a poor estimate. Significant advances have been made in the last few years on optimal estimation of a high-dimensional covariance matrix. Regularization methods such as banding, tapering, and thresholding have been proposed. In particular, Cai et al. (2010) established the optimal rate of convergence and Cai and Yuan (2012) developed an adaptive estimator of bandable covariance matrices. For sparse covariance matrices where each row and each column has relatively few nonzero entries, Bickel and Levina (2008) introduced a thresholding estimator and obtained rates of convergence; Cai and Liu (2011) proposed an adaptive thresholding procedure and Cai and Zhou (2012) established the minimax rates of convergence for estimating sparse covariance matrices.

Structural assumptions on the individual correlation matrices \mathbf{R}_1 and \mathbf{R}_2 are crucial for the good performance of the difference estimator. These assumptions, however, may not hold in practice. For example, gene transcriptional networks often contain the so-called hub nodes where the corresponding gene expressions are correlated with many other gene expressions. See, for example, (Barabási and Oltvai, 2004; Barabási et al., 2011). In such settings, some of the rows and columns of \mathbf{R}_1 and \mathbf{R}_2 have many nonzero entries which mean that \mathbf{R}_1 and \mathbf{R}_2 are not sparse. In genomic applications, the correlation matrices are rarely bandable as the genes are not ordered in any particular way.

In this paper, we propose a direct estimation method for the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ without first estimating \mathbf{R}_1 and \mathbf{R}_2 individually. This direct estimation method assumes that \mathbf{D} is approximately sparse, but otherwise does not impose any structural assumptions on the individual correlation matrices \mathbf{R}_1 and \mathbf{R}_2 . An adaptive thresholding procedure is introduced and analyzed. The estimator can still perform well even when the individual correlation matrices cannot be estimated consistently. For example, direct estimation can recover the differential correlation network accurately even in the presence of hub nodes in \mathbf{R}_1 and \mathbf{R}_2 .

as long as the differential correlation network is approximately sparse. The key is that sparsity is assumed for \mathbf{D} and not for \mathbf{R}_1 or \mathbf{R}_2 .

Theoretical performance guarantees are provided for direct estimator of the differential correlation matrix. Minimax rates of convergence are established for the collections of paired correlation matrices with approximately sparse differences. The proposed estimator is shown to be adaptively rate-optimal. In comparison to adaptive estimation of a single sparse covariance matrix considered in Cai and Liu (2011), both the procedure and the technical analysis of our method are different and more involved. Technically speaking, correlation matrix estimators are harder to analyze than those of covariance matrices and the two-sample setting in our problem further increases the difficulty.

Numerical performance of the proposed estimator is investigated through simulations. The results indicate significant advantage of estimating the differential correlation matrix directly. The estimator outperforms two other natural alternatives that are based on separate estimation of \mathbf{R}_1 and \mathbf{R}_2 . To further illustrate the merit of the method, we apply the procedure to the analysis of a breast cancer dataset from the study by van de Vijver et al. (2002) and investigate the differential co-expressions among genes in different tumor stages of breast cancer. The adaptive thresholding procedure is applied to analyze the difference in the correlation alternation in different grades of tumor. The study provides evidence at the gene co-expression level that several genes, of which a subset has been previously verified, are associated with the breast cancer.

In addition to optimal estimation of the differential correlation matrix, we also consider hypothesis testing of the differential correlation matrices, $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$ versus $H_1 : \mathbf{R}_1 - \mathbf{R}_2 \neq 0$. We propose a test which is particularly well suited for testing again sparse alternatives. The same ideas and techniques can also be used to treat other related problems. We also consider estimation of a single sparse correlation matrix from one random sample, estimation of the differential covariance matrices as well as estimation of the differential cross-correlation matrices.

The rest of the paper is organized as follows. Section 2 presents in detail the adaptive thresholding procedure for estimating the differential correlation matrix. The theoretical properties of the proposed estimator are analyzed in Section 3. In Section 4, simulation studies are carried out to investigate the numerical performance of the thresholding estimator and Section 5 illustrates the procedure through an analysis of a breast cancer dataset. Hypothesis testing on

the differential correlation matrices is discussed in Section 6.1, and other related problems are considered in the rest of Section 6. All the proofs are given in the Appendix.

2 Estimation of Differential Correlation Matrix

We consider in this section estimation of the differential correlation matrix and introduce a data-driven adaptive thresholding estimator. The theoretical and numerical properties of the estimator are investigated in Sections 3 and 4 respectively.

Let $\mathbf{X}^{(t)} = (X_1^{(t)}, \dots, X_p^{(t)})^\top$ be a p -variate random vector with mean $\boldsymbol{\mu}_t$, covariance matrix $\boldsymbol{\Sigma}_t = (\sigma_{ijt})_{1 \leq i, j \leq p}$, and correlation matrix $\mathbf{R}_t = (r_{ijt})_{1 \leq i, j \leq p}$, for $t = 1$ and 2 . Suppose we observe two i.i.d. random samples, $\{\mathbf{X}_1^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}\}$ from $\mathbf{X}^{(1)}$ and $\{\mathbf{X}_1^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}\}$ from $\mathbf{X}^{(2)}$, and the two samples are independent. The goal is to estimate the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ under the assumption that \mathbf{D} is approximately sparse.

Given the two random samples, the sample covariance matrices and sample correlation matrices are defined as, for $t = 1$ and 2 ,

$$\hat{\boldsymbol{\Sigma}}_t = (\hat{\sigma}_{ijt})_{1 \leq i, j \leq p} = \frac{1}{n_t} \sum_{k=1}^{n_t} (\mathbf{X}_k^{(t)} - \bar{\mathbf{X}}^{(t)})(\mathbf{X}_k^{(t)} - \bar{\mathbf{X}}^{(t)})^\top, \quad (1)$$

$$\hat{\mathbf{R}}_t = (\hat{r}_{ijt})_{1 \leq i, j \leq p} = \text{diag}(\hat{\boldsymbol{\Sigma}}_t)^{-1/2} \cdot \hat{\boldsymbol{\Sigma}}_t \cdot \text{diag}(\hat{\boldsymbol{\Sigma}}_t)^{-1/2}, \quad (2)$$

where $\bar{\mathbf{X}}^{(t)} = \frac{1}{n_t} \sum_{k=1}^{n_t} \mathbf{X}_k^{(t)}$ and $\text{diag}(\hat{\boldsymbol{\Sigma}}_t)$ is the diagonal matrix with the same diagonal as $\hat{\boldsymbol{\Sigma}}_t$. We propose a thresholding estimator of the differential correlation matrix \mathbf{D} by individually thresholding the entries of the difference of the two sample correlation matrices $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ with the threshold adaptive to the noise level of each entry. A key to the construction of the procedure is the estimation of the noise levels of the individual entries of $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$, as these entries are random variables themselves.

We first provide some intuition before formally introducing the estimate of the noise levels of the individual entries of $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$. Note that $E((X_i^{(t)} - \mu_{it})(X_j^{(t)} - \mu_{jt})) = \sigma_{ijt}$ and $\mu_{it} \approx \bar{X}_i^{(t)} = \frac{1}{n_t} \sum_{k=1}^{n_t} X_{ik}$. Define

$$\theta_{ijt} = \text{var}((X_i^{(t)} - \mu_{it})(X_j^{(t)} - \mu_{jt})), \quad 1 \leq i, j \leq p, \quad t = 1, 2. \quad (3)$$

Then one can intuitively write

$$\hat{\sigma}_{ijt} = \frac{1}{n_t} \sum_{k=1}^{n_t} (X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)}) \approx \sigma_{ijt} + \left(\frac{\theta_{ijt}}{n_t}\right)^{1/2} z_{ijt}, \quad (4)$$

where z_{ijt} is approximately normal with mean 0 and variance 1. Hence, θ_{ijt}/n_t measures the uncertainty of the sample covariance $\hat{\sigma}_{ijt}$. Based on the first order Taylor expansion of the 3-variate function $\frac{x}{(yz)^{1/2}}$ for $x \in \mathbb{R}$, and $y, z > 0$,

$$\frac{\hat{x}}{(\hat{y}\hat{z})^{1/2}} = \frac{x}{(yz)^{1/2}} + \frac{\hat{x} - x}{(yz)^{1/2}} - \frac{x}{(yz)^{1/2}} \left(\frac{\hat{y} - y}{2y} + \frac{\hat{z} - z}{2z} \right) + o(\hat{x} - x) + o(\hat{y} - y) + o(\hat{z} - z), \quad (5)$$

the entries \hat{r}_{ijt} of the sample correlation matrix $\hat{\mathbf{R}}_t = (\hat{r}_{ijt})$ can be approximated by

$$\begin{aligned} \hat{r}_{ijt} &= \frac{\hat{\sigma}_{ijt}}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} \approx \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} + \left(\frac{\theta_{ijt}}{n_t\sigma_{iit}\sigma_{jjt}} \right)^{1/2} z_{ijt} \\ &\quad - \frac{\sigma_{ijt}}{2(\sigma_{iit}\sigma_{jjt})^{1/2}} \left(\left(\frac{\theta_{iit}}{n_t\sigma_{iit}^2} \right)^{1/2} z_{iit} + \left(\frac{\theta_{jjt}}{n_t\sigma_{jjt}^2} \right)^{1/2} z_{jjt} \right) \\ &= r_{ijt} + \left(\frac{\xi_{ijt}}{n_t} \right)^{1/2} z_{ijt} - \frac{r_{ijt}}{2} \left(\left(\frac{\xi_{iit}}{n_t} \right)^{1/2} z_{iit} + \left(\frac{\xi_{jjt}}{n_t} \right)^{1/2} z_{jjt} \right), \end{aligned} \quad (6)$$

where we denote

$$\xi_{ijt} = \frac{\theta_{ijt}}{\sigma_{iit}\sigma_{jjt}}, \quad 1 \leq i, j \leq p, \quad t = 1, 2.$$

It then follows from (6) that

$$\begin{aligned} \hat{r}_{ij1} - \hat{r}_{ij2} &\approx r_{ij1} - r_{ij2} + \left(\frac{\xi_{ij1}}{n_1} \right)^{1/2} z_{ij1} - \frac{r_{ij1}}{2} \left(\left(\frac{\xi_{i11}}{n_1} \right)^{1/2} z_{i11} + \left(\frac{\xi_{j11}}{n_1} \right)^{1/2} z_{j11} \right) \\ &\quad - \left(\left(\frac{\xi_{ij2}}{n_2} \right)^{1/2} z_{ij2} - \frac{r_{ij2}}{2} \left(\left(\frac{\xi_{i22}}{n_2} \right)^{1/2} z_{i22} + \left(\frac{\xi_{j22}}{n_2} \right)^{1/2} z_{j22} \right) \right), \quad 1 \leq i, j \leq p, \end{aligned} \quad (7)$$

where the random variables z_{ij1} and z_{ij2} are approximately normal with mean 0 and variance 1, but not necessarily independent for $1 \leq i, j \leq p$.

Equation (7) suggests that estimation of $r_{ij1} - r_{ij2}$ is similar to the sparse covariance matrix estimation considered in Cai and Liu (2011), where it is proposed to adaptively threshold entries according to their individual noise levels. However, the setting here is more complicated as $\hat{r}_{ij1} - \hat{r}_{ij2}$ is not an unbiased estimate of $r_{ij1} - r_{ij2}$ and the noise levels are harder to estimate. These make the technical analysis more involved. The noise levels are unknown here but can be estimated based on the observed data. Specifically, we estimate θ_{ijt} and ξ_{ijt} by the following data-driven quantities,

$$\hat{\theta}_{ijt} = \frac{1}{n_t} \sum_{k=1}^{n_t} \left((X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)}) - \hat{\sigma}_{ijt} \right)^2, \quad (8)$$

$$\hat{\xi}_{ijt} = \frac{\hat{\theta}_{ijt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jjt}} = \frac{1}{n_t\hat{\sigma}_{iit}\hat{\sigma}_{jjt}} \sum_{k=1}^{n_t} \left((X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)}) - \hat{\sigma}_{ijt} \right)^2. \quad (9)$$

We are now ready to introduce the adaptive thresholding estimator of $\mathbf{R}_1 - \mathbf{R}_2$ using data-driven threshold levels. Let $s_\lambda(z)$ be a thresholding function satisfying the following conditions:

(C1). $|s_\lambda(z)| \leq c|y|$ for all z, y satisfying $|z - y| \leq \lambda$ for some $c > 0$;

(C2). $s_\lambda(z) = 0$ for $|z| \leq \lambda$;

(C3). $|s_\lambda(z) - z| \leq \lambda$, for all $z \in \mathbb{R}$.

Note that the commonly used soft thresholding function $s_\lambda(z) = \text{sgn}(z)(z - \lambda)_+$ and the adaptive lasso rule $s_\lambda = z(1 - |\lambda/z|^\eta)_+$ with $\eta \geq 1$ satisfy these three conditions. See Rothman et al. (2009) and Cai and Liu (2011). Although the hard thresholding function $s_\lambda(z) = z \cdot 1_{\{|z| \geq \lambda\}}$ does not satisfy Condition (C1), the technical arguments given in this paper still work with very minor changes.

We propose to estimate the sparse differential correlation matrix \mathbf{D} by the entrywise thresholding estimator $\hat{\mathbf{D}}^* = (\hat{d}_{ij}^*) \in \mathbb{R}^{p \times p}$ defined as

$$\hat{d}_{ij}^* = s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}), \quad 1 \leq i, j \leq p,$$

where $s_\lambda(z)$ is a thresholding function satisfying (C1)-(C3) and the threshold level λ_{ij} is given by $\lambda_{ij} = \lambda_{ij1} + \lambda_{ij2}$ with

$$\lambda_{ijt} = \tau \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|\hat{r}_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjet}^{1/2} \right) \right), \quad 1 \leq i, j \leq p, \quad t = 1, 2. \quad (10)$$

Here $\hat{\xi}_{ijt}$ are given by (9) and the thresholding constant τ can be chosen empirically through cross-validation. See Section 4.1 for more discussions on the empirical choice of τ .

3 Theoretical Properties

We now analyze the theoretical properties of the data-driven thresholding estimator $\hat{\mathbf{D}}^*$ proposed in the last section. We will establish the minimax rate of convergence for estimating the differential correlation matrix \mathbf{D} over certain classes of paired correlation matrices $(\mathbf{R}_1, \mathbf{R}_2)$ with approximately sparse difference $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ under the spectral norm loss. The results show that $\hat{\mathbf{D}}^*$ is rate-optimal under mild conditions.

3.1 Rate Optimality of the Thresholding Estimator

We consider the following class of paired correlation matrices in $\mathbb{R}^{p \times p}$ with approximately sparse difference

$$\mathcal{G}_q(s_0(p)) = \left\{ (\mathbf{R}_1, \mathbf{R}_2) : \mathbf{R}_1, \mathbf{R}_2 \succeq 0; \text{diag}(\mathbf{R}_1) = \text{diag}(\mathbf{R}_2) = 1; \max_i \sum_j |r_{ij1} - r_{ij2}|^q \leq s_0(p) \right\} \quad (11)$$

for some $0 \leq q < 1$. Here $\mathbf{R}_1, \mathbf{R}_2 \succeq 0$ and $\text{diag}(\mathbf{R}_1) = \text{diag}(\mathbf{R}_2) = 1$ mean that \mathbf{R}_1 and \mathbf{R}_2 are symmetric, semi-positive definite, and with all diagonal entries 1. For $(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))$, their difference $\mathbf{R}_1 - \mathbf{R}_2$ is approximately sparse in the sense that each row vector of $\mathbf{R}_1 - \mathbf{R}_2$ lies in the ℓ_q ball with radius $s_0(p)$ and $0 \leq q < 1$. When $q = 0$, this constraint becomes the commonly used exact sparsity condition.

Let

$$Y_i^{(t)} = (X_i^{(t)} - \mu_{it}) / (\text{var}(X_i^{(t)}))^{1/2}, \quad i = 1, \dots, p, \quad t = 1, 2.$$

We assume that for each i , Y_i is sub-Gaussian distributed, i.e. there exist constants $K, \eta > 0$ such that for all $1 \leq i \leq p$ and $t = 1, 2$,

$$E e^{u(Y_i^{(t)})^2} \leq K, \quad \text{for } |u| \leq \eta. \quad (12)$$

In addition, we assume for some constant $\nu_0 > 0$

$$\min_{1 \leq i, j \leq p; t=1,2} \text{var}(Y_i^{(t)} Y_j^{(t)}) \geq \nu_0. \quad (13)$$

The following theorem provides an upper bound for the risk of the thresholding estimator $\hat{\mathbf{D}}^*$ under the spectral norm loss.

Theorem 3.1 (Upper bound) *Suppose $\log p = o(\min(n_1, n_2)^{1/3})$ and (12) and (13) hold. Suppose the thresholding function $s_\lambda(z)$ satisfy Conditions (C1)-(C3). Then the thresholding estimator $\hat{\mathbf{D}}^*$ defined in (2) and (10) with $\tau > 4$ satisfies*

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (14)$$

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (15)$$

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 \leq Cp(s_0(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q/2} \quad (16)$$

for some constant $C > 0$ that does not depend on n_1, n_2 or p .

Remark 3.1 Condition (13) holds naturally when $\mathbf{X}^{(t)}$ are jointly Gaussian. To see this point, we suppose ρ_{ijt} is the correlation between $Y_i^{(t)}$ and $Y_j^{(t)}$. Then one can write $Y_j^{(t)} = \rho_{ijt}Y_i^{(t)} + \sqrt{1 - \rho_{ijt}^2}W$, where $Y_i^{(t)}, W$ are independently standard Gaussian. It is easy to calculate that $\text{Var}(Y_i^{(t)}Y_j^{(t)}) = 1 + \rho_{ijt}^2 \geq 1$, which implies (13) holds for $\nu_0 = 1$. Condition (13) is used in Lemma 6.1 to show that $\hat{\theta}_{ijt}$ is a good estimate of θ_{ijt} and $|\hat{\sigma}_{ijt} - \sigma_{ijt}|$ can be controlled by $C(\hat{\theta}_{ijt} \log p/n_t)^{1/2}$ with high probability.

Theorem 3.1 gives the rate of convergence for the thresholding estimator $\hat{\mathbf{D}}^*$. The following result provides the lower bound for the minimax risk of estimating the differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ with $(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))$.

Theorem 3.2 (Lower Bound) *Suppose $\log p = o(\min(n_1, n_2))$ and $s_0(p) \leq M \min(n_1, n_2)^{(1-q)/2} \times (\log p)^{-(3-q)/2}$ for some constant $M > 0$. Then minimax risk for estimating $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ satisfies*

$$\inf_{\hat{\mathbf{D}}} \sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}} - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \geq cs_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}, \quad (17)$$

$$\inf_{\hat{\mathbf{D}}} \sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}} - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 \geq cs_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}, \quad (18)$$

$$\inf_{\hat{\mathbf{D}}} \sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}} - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 \geq cs_0(p)p \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q/2}, \quad (19)$$

for some constant $c > 0$.

Theorems 3.1 and 3.2 together yield the minimax rate of convergence

$$s_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}$$

for estimating $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ with $(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))$ under the spectral norm loss, and show that the thresholding estimator $\hat{\mathbf{D}}^*$ defined in (2) and (10) is adaptively rate-optimal.

Remark 3.2 The technical analysis here for the different of two correlation matrices is more complicated in comparison to the problem of estimating a sparse covariance matrix considered in Cai and Liu (2011). It can be seen in (7), i.e. the “signal + noise” expression of $\hat{r}_{ij1} - \hat{r}_{ij2}$, the difference of the sample correlation matrices has six “noise terms”. It is necessary to deal with all these six terms in the theoretical analysis of Theorem 3.1.

4 Numerical Studies

We investigate in this section the numerical performance of the adaptive thresholding estimator of the differential correlation matrix through simulations. The method is applied to the analysis of a breast cancer dataset in the next section.

In the previous sections, we proposed the entrywise thresholding method for estimating $\mathbf{R}_1 - \mathbf{R}_2$ and then studied the theoretical properties of $\hat{\mathbf{D}}^*$ with a fixed $\tau > 4$. However, the theoretical choice of τ may not be optimal in finite sample performance, as we can see in the following example. Let \mathbf{R}_1 and \mathbf{R}_2 be 200×200 -dimensional matrices such that $\mathbf{R}_{1,ij} = (-1)^{|i-j|} \times \max(1 - |i-j|/10, 0) \times (1_{\{i=j\}} + f_i f_j 1_{\{i \neq j\}})$ and $\mathbf{R}_{2,ij} = \max(1 - |i-j|/10, 0) \times (1_{\{i=j\}} + f_i f_j 1_{\{i \neq j\}})$. Here $1_{\{\cdot\}}$ is the indicator function, f_1, \dots, f_{200} are i.i.d. random variables that are uniformly distributed on $[0, 1]$. In this setting, both \mathbf{R}_1 and \mathbf{R}_2 are sparse, but their difference is even more sparse. We set $\Sigma_t = \mathbf{R}_t$ and generate 200 independent samples from $\mathbf{X}^{(1)} \sim N(0, \Sigma_1)$ and 200 independent samples from $\mathbf{X}^{(2)} \sim N(0, \Sigma_2)$. For various values of $\tau \in [0, 5]$, we implement the proposed method with hard thresholding and repeat the experiments for 100 times. The average loss in spectral, ℓ_1 and Frobenius norms are shown in Figure 1. Obviously

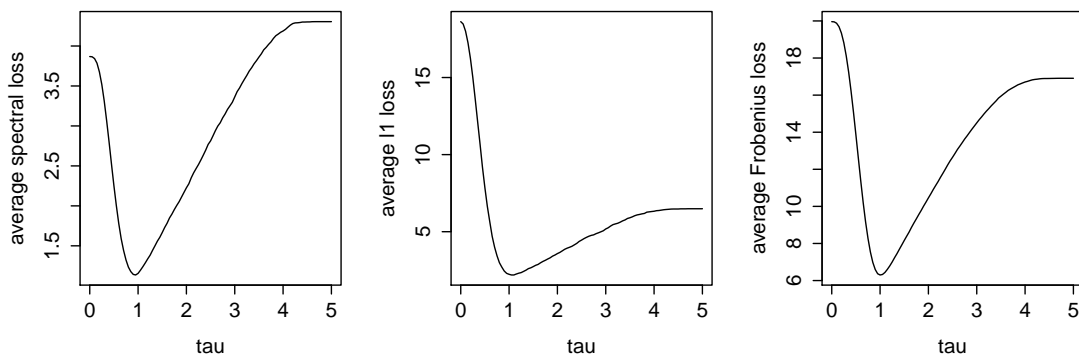


Figure 1: Average (Spectral, ℓ_1 , Frobenius) norm losses for $\tau \in [0, 5]$. $p = 100$, $n_1 = n_2 = 50$.

in this example, $\tau > 4$ is not the best choice.

Empirically, we find that the numerical performance of the estimator can often be improved by using a data-driven choice of τ based on cross-validation. We thus begin by introducing the following K -fold cross-validation method for the empirical selection of τ .

4.1 Empirical Choice of τ

For an integer $K \geq 2$, we first divide both samples $\mathbf{X}^{(1)} = \{\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \dots, \mathbf{X}_{n_1}^{(1)}\}$ and $\mathbf{X}^{(2)} = \{\mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \dots, \mathbf{X}_{n_2}^{(2)}\}$ randomly into two groups for H times as $\mathbf{X}_{I_1^h}^{(1)}, \mathbf{X}_{T_1^h}^{(1)}, \mathbf{X}_{I_2^h}^{(2)}$ and $\mathbf{X}_{T_2^h}^{(2)}$. Here $h = 1, \dots, H$ represents the h -th division. For $t = 1$ and 2 , the size of the first group $\mathbf{X}_{I_t^h}^{(t)}$ is approximately $(K-1)/K \cdot n_t$ and the size of the second group $\mathbf{X}_{T_t^h}^{(t)}$ is approximately n_t/K . We then calculate the corresponding sample correlation matrices as $\hat{\mathbf{R}}_{I_1^h}^{(1)}, \hat{\mathbf{R}}_{T_1^h}^{(1)}, \hat{\mathbf{R}}_{I_2^h}^{(2)}$ and $\hat{\mathbf{R}}_{T_2^h}^{(2)}$ for all four sub-samples. Partition the interval $[0, 5]$ into an equi-spaced grid $\{0, \frac{1}{N}, \dots, \frac{5N}{N}\}$. For each value of $\tau \in \{0, \frac{1}{N}, \dots, \frac{5N}{N}\}$, we obtain the thresholding estimator $\hat{\mathbf{D}}_{I^h}^*$ defined in (2) and (10) with the thresholding constant τ based on the subsamples $\mathbf{X}_{I_1^h}^{(1)}$ and $\mathbf{X}_{I_2^h}^{(2)}$. Denote the average loss for each τ for the second sub-samples $\mathbf{X}_{T_1^h}^{(1)}$ and $\mathbf{X}_{T_2^h}^{(2)}$ as

$$L(\tau) = \frac{1}{H} \sum_{h=1}^H \|\hat{\mathbf{D}}_{I^h}^* - (\hat{\mathbf{R}}_{T_1^h}^{(1)} - \hat{\mathbf{R}}_{T_2^h}^{(2)})\|_F^2.$$

We select

$$\hat{\tau} = \underset{\tau \in \{0, \frac{1}{N}, \dots, \frac{5N}{N}\}}{\operatorname{argmin}} L(\tau)$$

as our empirical choice of the thresholding constant τ , and calculate the final estimator $\hat{\mathbf{D}}^*(\hat{\tau})$ with the thresholding constant $\hat{\tau}$ based on the whole samples $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

4.2 Estimation of Differential Correlation Matrix

The adaptive thresholding estimator is easy to implement. We consider the following two models under which the differential correlation matrix is sparse.

1. Model 1 (Random Sparse Difference) \mathbf{R}_1 and \mathbf{R}_2 are p -dimensional symmetric positive definite matrices such that $\mathbf{R}_1 = \operatorname{diag}(\mathbf{B}_1, \mathbf{I}_{\frac{p}{2} \times \frac{p}{2}})$ is a fixed matrix, where $\mathbf{B}_1 \in \mathbb{R}^{\frac{p}{2} \times \frac{p}{2}}$ with $\mathbf{B}_{1,ij} = 1$ if $i = j$ and $\mathbf{B}_{1,ij} = 0.2$ if $i \neq j$, $\mathbf{I}_{\frac{p}{2} \times \frac{p}{2}}$ is the $\frac{p}{2} \times \frac{p}{2}$ identity matrix, and \mathbf{R}_2 is randomly generated as $\mathbf{R}_2 = \operatorname{diag}(\mathbf{B}_1 + \lambda \mathbf{D}_0, \mathbf{I}_{\frac{p}{2} \times \frac{p}{2}})$, where $\mathbf{D}_0 \in \mathbb{R}^{\frac{p}{2} \times \frac{p}{2}}$ with

$$\mathbf{D}_{ij,0} = \begin{cases} 1, & \text{with probability } 0.05 \\ 0, & \text{with probability } 0.9 \\ -1, & \text{with probability } 0.05 \end{cases}$$

and λ is a constant that ensures the positive definiteness of \mathbf{R}_2 .

2. Model 2 (Banded Difference) In this setting, p -dimensional matrices \mathbf{R}_1 and \mathbf{R}_2 satisfy $\mathbf{R}_{1,ij} = 0.2 \times 1_{\{i=j\}} + 0.8 \times (-1)^{|i-j|} \times \max(1 - |i-j|/10, 0)$ and $\mathbf{R}_{2,ij} = \mathbf{R}_{1,ij} + 0.2 \times 1_{\{i \neq j\}} \times \max(1 - |i-j|/3, 0)$. Here $1_{\{\cdot\}}$ is the indicator function.

In each of the two settings, we set $\boldsymbol{\Sigma}_t = \text{diag}(|\omega_t|^{1/2}) \mathbf{R}_t \text{diag}(|\omega_t|^{1/2})$ for both $t = 1, 2$, where $\omega_1, \omega_2 \in \mathbb{R}^p$ are two i.i.d. samples from $N(0, \mathbf{I}_p)$. These operations make the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ have different values along the diagonals.

We generate i.i.d. samples from $\mathbf{X}^{(1)} \sim N(0, \boldsymbol{\Sigma}_1)$ and $\mathbf{X}^{(2)} \sim N(0, \boldsymbol{\Sigma}_2)$ for various values of p, n_1 , and n_2 and then apply the proposed algorithm with 5-fold cross-validation for the selection of the thresholding constant τ . For each setting, both the hard thresholding and adaptive-Lasso thresholding (Rothman et al. (2009)),

$$s_\lambda(z) = z \cdot \max(1 - |\lambda/z|^\eta, 0) \quad \text{with} \quad \eta = 4, \quad (20)$$

are used. For comparison, we also implement three natural estimators of \mathbf{D} .

1. The covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are estimated individually by the adaptive thresholding method proposed in Cai and Liu (2011) with 5-fold cross-validation and then $\hat{\boldsymbol{\Sigma}}_1^*$ and $\hat{\boldsymbol{\Sigma}}_2^*$ are normalized to yield estimators of \mathbf{R}_1 and \mathbf{R}_2 ,

$$\hat{\mathbf{R}}_1^* = \text{diag}(\hat{\boldsymbol{\Sigma}}_1^*)^{-1/2} \hat{\boldsymbol{\Sigma}}_1^* \text{diag}(\hat{\boldsymbol{\Sigma}}_1^*)^{-1/2}, \quad \hat{\mathbf{R}}_2^* = \text{diag}(\hat{\boldsymbol{\Sigma}}_2^*)^{-1/2} \hat{\boldsymbol{\Sigma}}_2^* \text{diag}(\hat{\boldsymbol{\Sigma}}_2^*)^{-1/2},$$

and finally $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$ is estimated by the difference $\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$.

2. The correlation matrices $\hat{\mathbf{R}}_1^\bullet$ and $\hat{\mathbf{R}}_2^\bullet$ are estimated separately using the method proposed in Section 6.2 and then take the difference.
3. \mathbf{D} is estimated directly the difference of the sample correlation matrices $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$.

The numerical results are summarized in Tables 1 and 4.2 for the two models respectively. In each case, we compare the performance of the three estimators \mathbf{D}^* , $\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$ and $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ under the spectral norm, matrix ℓ_1 norm, and Frobenius norm losses. For both models, it is easy to see that the direct thresholding estimator \mathbf{D}^* significantly outperforms $\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$ and $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$. Under Model 1, the individual correlation matrices \mathbf{R}_1 and \mathbf{R}_2 are “dense” in the sense that half of the rows and columns contain many non zeros entries, but their difference \mathbf{D} is sparse. In this case, \mathbf{R}_1 and \mathbf{R}_2 cannot be estimated consistently and the two difference estimators

$\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$ and $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ based on the individual estimators of \mathbf{R}_1 and \mathbf{R}_2 perform very poorly, while the direct estimator \mathbf{D}^* performs very well. Moreover, the numerical performance of the thresholding estimators does not depend on the specific thresholding rules in a significant way. Different thresholding rules including hard thresholding and adaptive Lasso behave similarly.

5 Analysis of A Breast Cancer Dataset

Identifying gene expression networks can be helpful for conducting more effective treatment based to the condition of patients. de la Fuente (2010) demonstrated that the gene expression networks can vary in different disease states and the differential correlations in gene expression (i.e. co-expression) are useful in disease studies.

In this section, we consider the dataset “70pathwaygenes-by-grade” from the study by van de Vijver et al. (2002) and investigate the differential co-expressions among genes in different tumor stages of breast cancer. In this dataset, there are 295 records of patients with 1624 gene expressions, which are categorized into three groups based on the histological grades of tumor (“Good”, “Intermediate” and “Poor”) with 74, 101 and 119 records, respectively. We denote these three groups of samples as $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$. In order to analyze the difference in the correlation alternation in different grades of tumor, we apply our adaptive thresholding method with cross-validation to estimate the differential correlation matrices among those gene expressions from different stages.

The number of gene pairs with significant difference in correlation are listed in Table 3. The results show that the correlation structures between the “Good” and “Intermediate” groups are similar and there is some significant changes between the “Good” and “Poor” group.

More interestingly, by combining the “Good” and “Intermediate” sub-samples and comparing with the “Poor” group, we find significant differences between their correlation structure. There are 4526 pairs of genes that have significantly different correlations between the “Good + Intermediate” and “Poor” groups. For each given gene, we count the number of the genes whose correlation with this gene is significantly different between these two groups, and rank all the genes by the counts. That is, we rank the genes by the size of the support of $\hat{\mathbf{D}}^*$ in each row. The top ten genes are listed in Table 4.

Among these ten genes, six of them, GDF5, TCF7L1, PAPSS1, SFRP1, GABRP, TGFB1,

p	n_1	n_2	Hard			Adaptive Lasso			Sample
			$\hat{\mathbf{D}}^*$	$\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$	$\hat{\mathbf{R}}_1^\bullet - \hat{\mathbf{R}}_2^\bullet$	$\hat{\mathbf{D}}^*$	$\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$	$\hat{\mathbf{R}}_1^\bullet - \hat{\mathbf{R}}_2^\bullet$	$\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$
Spectral Norm									
100	50	50	0.50(0.41)	1.75(1.37)	6.94(1.07)	0.33(0.31)	1.51(1.75)	6.17(0.98)	7.28(0.93)
100	100	100	0.34(0.21)	3.79(3.17)	4.74(0.55)	0.28(0.23)	3.53(2.71)	4.49(0.71)	5.02(0.65)
100	200	200	0.29(0.19)	4.22(2.14)	3.23(0.47)	0.24(0.13)	4.52(1.86)	3.14(0.54)	3.55(0.47)
100	500	500	0.24(0.10)	1.72(0.35)	2.07(0.32)	0.22(0.08)	1.82(0.35)	1.87(0.26)	2.23(0.25)
500	50	50	0.56(0.77)	3.02(2.76)	31.88(4.04)	0.40(0.65)	3.47(4.63)	29.15(4.26)	34.66(3.84)
500	100	100	0.41(0.48)	8.09(11.02)	23.38(4.52)	0.34(0.39)	12.99(13.26)	21.82(3.23)	24.19(2.77)
500	200	200	0.32(0.40)	22.22(13.06)	15.67(3.30)	0.26(0.34)	21.31(9.29)	14.61(2.24)	16.50(1.97)
500	500	500	0.20(0.19)	7.80(1.37)	7.80(1.29)	0.18(0.14)	8.21(1.69)	8.70(1.39)	10.46(1.31)
Matrix ℓ_1 Norm									
100	50	50	0.89(0.68)	3.63(2.97)	18.91(1.55)	0.78(0.80)	3.14(3.20)	16.88(1.42)	21.33(1.61)
100	100	100	0.64(0.25)	7.34(4.85)	13.42(1.08)	0.70(0.87)	7.03(4.14)	12.72(1.18)	14.97(1.14)
100	200	200	0.64(0.34)	9.63(1.85)	9.22(0.75)	0.61(0.37)	9.37(1.77)	8.67(0.87)	10.60(0.81)
100	500	500	0.58(0.22)	4.54(0.56)	5.92(0.54)	0.56(0.21)	4.85(0.61)	5.33(0.45)	6.69(0.44)
500	50	50	1.69(3.09)	7.85(8.14)	97.28(6.64)	1.37(2.87)	9.49(11.93)	87.02(7.31)	112.40(8.97)
500	100	100	1.06(1.28)	20.12(19.50)	64.98(5.93)	1.17(1.47)	27.95(22.75)	65.60(5.07)	79.66(5.37)
500	200	200	0.97(1.48)	51.64(10.13)	46.74(5.01)	0.95(1.24)	47.87(8.50)	45.54(3.70)	55.77(3.70)
500	500	500	0.68(0.54)	23.19(2.86)	23.47(2.56)	0.69(0.52)	24.67(2.92)	27.21(2.09)	35.32(2.04)
Frobenious Norm									
100	50	50	1.40(1.32)	4.34(2.51)	19.01(0.37)	1.06(1.04)	3.26(2.61)	16.60(0.42)	19.87(0.38)
100	100	100	0.96(0.59)	7.14(3.67)	13.38(0.23)	0.94(0.81)	6.23(3.31)	12.10(0.27)	14.05(0.25)
100	200	200	0.89(0.57)	9.09(1.03)	9.30(0.20)	0.84(0.50)	8.15(1.07)	8.42(0.25)	9.94(0.18)
100	500	500	0.85(0.32)	4.37(0.27)	5.92(0.11)	0.82(0.30)	4.43(0.25)	5.26(0.11)	6.39(0.10)
500	50	50	3.33(5.63)	11.18(7.71)	95.05(0.91)	2.27(3.92)	9.57(9.31)	83.40(1.24)	99.97(0.93)
500	100	100	2.18(2.98)	20.17(14.55)	61.37(2.09)	2.10(2.47)	22.54(16.69)	60.58(0.85)	70.40(0.67)
500	200	200	1.77(2.39)	45.06(5.89)	42.11(1.67)	1.63(1.96)	39.52(5.14)	41.88(0.73)	49.53(0.49)
500	500	500	1.27(1.09)	20.17(0.99)	21.74(0.65)	1.22(0.84)	20.48(0.91)	25.47(0.41)	31.34(0.33)

Table 1: Comparison of $\hat{\mathbf{D}}^*$ with $\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$ and $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ under Model 1.

p	n_1	n_2	Hard			Adaptive Lasso			Sample
			$\hat{\mathbf{D}}^*$	$\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$	$\hat{\mathbf{R}}_1^\bullet - \hat{\mathbf{R}}_2^\bullet$	$\hat{\mathbf{D}}^*$	$\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$	$\hat{\mathbf{R}}_1^\bullet - \hat{\mathbf{R}}_2^\bullet$	$\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$
Spectral Norm									
100	50	50	0.98(1.00)	4.61(1.49)	7.25(0.87)	0.71(0.70)	4.47(1.44)	6.05(0.74)	8.29(0.98)
100	100	100	0.70(0.51)	2.88(0.81)	5.01(0.57)	0.62(0.47)	2.93(0.87)	4.25(0.52)	5.83(0.59)
100	200	200	0.60(0.35)	1.93(0.55)	3.53(0.42)	0.48(0.24)	1.98(0.57)	2.98(0.37)	4.07(0.47)
100	500	500	0.47(0.14)	1.23(0.27)	2.32(0.23)	0.46(0.17)	1.30(0.36)	1.98(0.21)	2.66(0.27)
500	50	50	0.97(0.99)	5.03(1.00)	20.61(1.07)	0.80(0.75)	4.55(0.96)	16.91(0.90)	24.96(1.16)
500	100	100	0.79(0.62)	3.17(0.49)	13.64(0.59)	0.59(0.41)	3.14(0.63)	11.13(0.53)	16.39(0.71)
500	200	200	0.60(0.36)	2.13(0.30)	9.12(0.42)	0.51(0.31)	2.11(0.35)	7.44(0.37)	10.94(0.50)
500	500	500	0.51(0.20)	1.34(0.16)	5.63(0.23)	0.49(0.20)	1.35(0.22)	4.65(0.21)	6.78(0.29)
Matrix ℓ_1 Norm									
100	50	50	1.84(2.66)	10.61(3.48)	19.11(1.55)	1.26(1.86)	9.88(3.14)	16.18(1.55)	21.92(1.61)
100	100	100	1.18(1.44)	6.73(2.24)	13.62(1.08)	1.10(1.26)	6.73(2.12)	11.73(1.20)	15.87(1.11)
100	200	200	0.98(0.98)	4.53(1.47)	9.79(0.85)	0.71(0.71)	4.68(1.46)	8.39(0.89)	11.37(0.97)
100	500	500	0.67(0.48)	2.95(0.89)	6.44(0.56)	0.65(0.59)	3.08(1.03)	5.58(0.47)	7.47(0.53)
500	50	50	1.79(2.65)	11.03(2.80)	79.71(2.64)	1.64(2.26)	10.38(3.02)	64.46(2.73)	97.88(2.55)
500	100	100	1.45(1.75)	7.66(1.79)	56.52(2.16)	1.02(1.35)	7.73(2.40)	45.65(1.86)	69.42(1.76)
500	200	200	1.02(1.18)	4.97(1.09)	39.86(1.39)	0.83(1.14)	5.03(1.27)	31.90(1.15)	49.11(1.33)
500	500	500	0.81(0.70)	3.15(0.72)	25.34(0.77)	0.82(0.77)	3.27(1.00)	20.39(0.77)	31.36(0.79)
Frobenious Norm									
100	50	50	3.36(2.53)	13.82(1.83)	18.46(0.81)	2.66(1.47)	12.13(2.00)	15.87(0.79)	19.92(0.94)
100	100	100	2.67(1.19)	9.46(1.28)	13.26(0.55)	2.54(1.10)	8.77(1.18)	11.51(0.54)	14.32(0.58)
100	200	200	2.43(0.69)	6.94(0.72)	9.75(0.39)	2.26(0.51)	6.68(0.76)	8.59(0.37)	10.49(0.43)
100	500	500	2.24(0.34)	5.29(0.36)	6.96(0.19)	2.25(0.46)	5.28(0.44)	6.33(0.17)	7.39(0.19)
500	50	50	6.77(4.86)	34.24(3.33)	91.09(0.85)	6.18(3.86)	27.39(3.39)	75.97(0.83)	100.71(0.92)
500	100	500	6.19(2.98)	22.76(1.92)	64.37(0.56)	5.30(1.79)	20.12(2.21)	53.72(0.56)	71.23(0.58)
500	200	200	5.32(1.49)	16.34(1.18)	45.79(0.44)	5.10(1.36)	15.01(1.15)	38.36(0.42)	50.61(0.43)
500	500	500	5.00(0.62)	12.14(0.59)	29.77(0.27)	4.99(0.69)	11.76(0.70)	25.27(0.24)	32.80(0.25)

Table 2: Comparison of $\hat{\mathbf{D}}^*$ with $\hat{\mathbf{R}}_1^* - \hat{\mathbf{R}}_2^*$ and $\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2$ under Model 2.

	Good v.s. Intermediate	Intermediate v.s. Poor	Good v.s. Poor
# of selected pairs	0	2	152

Table 3: The number of gene pairs that have significant differential correlation between two groups of different tumor grades

Gene	number of pairs
growth differentiation factor 5 (GDF5)	67
transcription factor 7-like 1 (TCF7L1)	64
3'-phosphoadenosine 5'-phosphosulfate synthase 1 (PAPSS1)	51
secreted frizzled-related protein 1(SFRP1)	43
gamma-aminobutyric acid A receptor, pi (GABRP)	41
mannosidase, alpha, class 2B, member 2 (MAN2B2)	37
desmocollin 2 (DSC2)	36
transforming growth factor, beta 3 (TGFB3)	35
CRADD	35
ELOVL fatty acid elongase 5(ELOVL5)	32

Table 4: The top ten genes that appear for most times in the selected pairs in “Good + Intermediate” v.s. “Poor”

have been previously studied and verified in the literature that are associated with the breast cancer (See Margheri et al. (2012), Shy et al. (2013), Xu et al. (2012), Klopocki et al. (2004), Zafrakas et al. (2006), and Ghellal et al. (2000), respectively). Take for example GDF5 and TCF7L1, the overproduction of Transforming growth factor beta-1 ($TGF\beta$), a multifunctional cytokine, is an important characteristic of late tumor progression. Based on the study by Margheri et al. (2012), $TGF\beta$ produced by breast cancer cells brings about in endothelial cells expression of GDF5. The findings in (Shy et al. (2013)) suggested the important role played by TCF7L1 in breast cancer. Although these biological studies mainly focus on the the behavior of the single gene expression, our study provides evidence in the gene co-expression level that these gene expressions are related with the breast cancer.

We should point out that the two well-known genes related to the breast cancer, BRCA1

and BRCA2, were not detected by our method. This is mainly due to the fact that our method focus on the differential gene co-expressions, not the changes in the gene expression levels.

6 Other Related Problems

We have so far focused on optimal estimation of the differential correlation matrix. In addition to optimal estimation, hypothesis testing of the differential correlation matrix is also an important problem. In this section we consider testing the hypotheses $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$ versus $H_1 : \mathbf{R}_1 - \mathbf{R}_2 \neq 0$ and propose a test which is particularly well suited for testing against sparse alternatives.

Similar ideas and techniques can also be used to treat several other related problems, including estimation of a single sparse correlation matrix from one random sample, estimation of the differential covariance matrices, and estimation of the differential cross-correlation matrices. We also briefly discuss these problems in this section.

6.1 Testing Differential Correlation Matrices

Suppose we are given two sets of independent and identical distributed samples $\mathbf{X}^{(t)} = \{\mathbf{X}_1^{(t)}, \dots, \mathbf{X}_{n_t}^{(t)}\}$ with the mean μ_t , covariance matrix $\mathbf{\Sigma}_t$ and correlation matrix \mathbf{R}_t , where $t = 1$ and 2 , and wish to test the hypotheses

$$H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0 \quad \text{v.s.} \quad H_1 : \mathbf{R}_1 - \mathbf{R}_2 \neq 0. \quad (21)$$

This testing problem is similar to, but also different from, testing the equality of two high-dimensional covariance matrices, which has been considered in several recent papers. See, for example, Schott (2007), Srivastava and Yanagihara (2010), Li et al. (2012), and Cai et al. (2013). Here we are particularly interested in testing against sparse alternatives and follow similar ideas as those in Cai et al. (2013).

To construct the test statistic, we need more precise understanding of the sample correlation

coefficients \hat{r}_{ijt} . It follows from (5) that

$$\begin{aligned}\hat{r}_{ijt} &= \frac{\hat{\sigma}_{ijt}}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} \approx \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} + \frac{\hat{\sigma}_{ijt} - \sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} - \frac{\sigma_{ijt}}{2(\sigma_{iit}\sigma_{jjt})^{1/2}} \left(\frac{\hat{\sigma}_{iit} - \sigma_{iit}}{(\sigma_{iit}\sigma_{iit})^{1/2}} + \frac{\hat{\sigma}_{jjt} - \sigma_{jjt}}{(\sigma_{jjt}\sigma_{jjt})^{1/2}} \right) \\ &= r_{ijt} + \frac{1}{n_t} \sum_{k=1}^{n_t} \left[\frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)}) - \sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} \right. \\ &\quad \left. - \frac{r_{ijt}}{2} \left(\frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})^2 - \sigma_{iit}}{\sigma_{iit}} + \frac{(X_{jk}^{(t)} - \bar{X}_j^{(t)})^2 - \sigma_{jjt}}{\sigma_{jjt}} \right) \right]\end{aligned}$$

Since $\bar{X}_i^{(t)} \approx \mu_{it}$, $\bar{X}_j^{(t)} \approx \mu_{jt}$, $E \left(X_{ik}^{(t)} - \bar{X}_i^{(t)} \right) \left(X_{jk}^{(t)} - \bar{X}_j^{(t)} \right) \approx \sigma_{ijt}$, we introduce

$$\eta_{ijt} = \text{var} \left[\frac{(X_i^{(t)} - \mu_{it})(X_j^{(t)} - \mu_{jt})}{(\sigma_{iit}\sigma_{jjt})^{1/2}} - \frac{r_{ijt}}{2} \left(\frac{(X_i^{(t)} - \mu_{it})^2}{\sigma_{iit}} + \frac{(X_j^{(t)} - \mu_{jt})^2}{\sigma_{jjt}} \right) \right].$$

Then asymptotically as $n, p \rightarrow \infty$,

$$\hat{r}_{ijt} - r_{ijt} \approx \left(\frac{\eta_{ijt}}{n_t} \right)^{1/2} z_{ijt}, \quad \text{where } z_{ijt} \sim N(0, 1).$$

The true value of η_{ijt} is unknown but can be estimated by

$$\begin{aligned}\hat{\eta}_{ijt} &= \frac{1}{n_t} \sum_{k=1}^{n_t} \left\{ \frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)}) - \hat{\sigma}_{ijt}}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} \right. \\ &\quad \left. - \frac{\hat{r}_{ijt}}{2} \left(\frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})^2 - \hat{\sigma}_{iit}}{\hat{\sigma}_{iit}} + \frac{(X_{jk}^{(t)} - \bar{X}_j^{(t)})^2 - \hat{\sigma}_{jjt}}{\hat{\sigma}_{jjt}} \right) \right\}^2 \\ &= \frac{1}{n_t} \sum_{k=1}^{n_t} \left\{ \frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})(X_{jk}^{(t)} - \bar{X}_j^{(t)})}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} - \frac{\hat{r}_{ijt}}{2} \left(\frac{(X_{ik}^{(t)} - \bar{X}_i^{(t)})^2}{\hat{\sigma}_{iit}} + \frac{(X_{jk}^{(t)} - \bar{X}_j^{(t)})^2}{\hat{\sigma}_{jjt}} \right) \right\}^2.\end{aligned}$$

We define the test statistic by

$$T_n = \max_{1 \leq i \leq j \leq p} T_{ij}$$

where

$$T_{ij} = \frac{(\hat{r}_{ij1} - \hat{r}_{ij2})^2}{\hat{\eta}_{ij1}/n_1 + \hat{\eta}_{ij2}/n_2}, \quad 1 \leq i, j \leq p.$$

Under regularity conditions (similar to (C1)-(C3) in Cai et al. (2013)), the asymptotic distribution of T_n can be shown to be the type I extreme value distribution. More precisely,

$$P(T_n - 4 \log p + \log \log p \leq t) \rightarrow \exp \left(-(8\pi)^{-1/2} \exp(-t/2) \right) \quad (22)$$

for any given $t \in \mathbb{R}$.

The asymptotic null distribution (22) can then be used to construct a test for testing the hypothesis $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$. For a given significance level $0 < \alpha < 1$, define the test Ψ_α by

$$\Psi_\alpha = I(M_n \geq 4 \log p - \log \log p + \tau_\alpha) \quad (23)$$

where $\tau_\alpha = -\log(8\pi) - 2 \log \log(1 - \alpha)^{-1}$ is the $1 - \alpha$ quantile of the type I extreme value distribution with the cumulative distribution function $\exp(-(8\pi)^{-1/2} \exp(-x/2))$. The hypothesis $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$ is rejected whenever $\Psi_\alpha = 1$. As the test proposed in Cai et al. (2013) for testing the equality of two covariance matrices, the test Ψ_α defined in (23) can also be shown to be particularly well suited for testing $H_0 : \mathbf{R}_1 - \mathbf{R}_2 = 0$ against sparse alternatives.

6.2 Optimal Estimation of a Sparse Correlation Matrix

The ideas and technical tools can also be used for estimation of a single correlation matrix from one random sample, which is a simpler problem. Suppose we observe an independent and identical distributed sample $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ from a p -dimensional distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$, covariance matrix $\boldsymbol{\Sigma}$, and correlation matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$. When \mathbf{R} is approximately sparse, it can be naturally estimated by a thresholding estimator $\hat{\mathbf{R}}$ as follows. Let $\bar{X}_i = \frac{1}{n} \sum_{k=1}^n X_{ik}$. Define the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}_{ij})_{1 \leq i, j \leq p}$ and the sample correlation matrix $\hat{\mathbf{R}} = (\hat{r}_{ij})_{1 \leq i, j \leq p}$ respectively by

$$\hat{\sigma}_{ij} = \frac{1}{n} \sum_{k=1}^n (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) \quad \text{and} \quad \hat{r}_{ij} = \frac{\hat{\sigma}_{ij}}{(\hat{\sigma}_{ii}\hat{\sigma}_{jj})^{1/2}}.$$

Same as in (8) and (9), we define

$$\hat{\theta}_{ij} = \frac{1}{n} \sum_{k=1}^n ((X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) - \hat{\sigma}_{ij})^2, \quad (24)$$

$$\hat{\xi}_{ij} = \frac{\hat{\theta}_{ij}}{\hat{\sigma}_{ii}\hat{\sigma}_{jj}} = \frac{1}{n\hat{\sigma}_{ii}\hat{\sigma}_{jj}} \sum_{k=1}^n ((X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j) - \hat{\sigma}_{ij})^2, \quad (25)$$

$$\lambda_{ij} = \tau \left(\frac{\log p}{n} \right)^{1/2} \left(\hat{\xi}_{ij}^{1/2} + \frac{|\hat{r}_{ij}|}{2} \left(\hat{\xi}_{ii}^{1/2} + \hat{\xi}_{jj}^{1/2} \right) \right), \quad (26)$$

where τ is the thresholding constant that can be chosen empirically through cross-validation.

The correlation matrix \mathbf{R} is then estimated by $\hat{\mathbf{R}}^* = (\hat{r}_{ij}^*)_{1 \leq i, j \leq p}$ with

$$\hat{r}_{ij}^* = s_{\lambda_{ij}}(\hat{r}_{ij}).$$

We consider the following class of approximately sparse correlation matrices

$$\mathcal{G}_q^1(s_0(p)) = \left\{ \mathbf{R} = (r_{ij}) : \mathbf{R} \succ 0; \text{diag}(\mathbf{R}) = 1; \max_j \sum_{i=1, i \neq j}^p |r_{ij}|^q \leq s_0(p) \right\}, \quad 0 \leq q < 1.$$

The following theoretical results for $\hat{\mathbf{R}}^*$ can be established using a similar analysis.

Proposition 6.1 *Suppose $\log p = o(n^{1/3})$ and \mathbf{X} satisfies (12), (13). For $\tau > 6$, there exists some constant C does not depend on n or p such that*

$$\sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}}^* - \mathbf{R}\|^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n} \right)^{1-q} \quad (27)$$

$$\sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}}^* - \mathbf{R}\|_{\ell_1}^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n} \right)^{1-q} \quad (28)$$

$$\sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}}^* - \mathbf{R}\|_F^2 \leq Cp(s_0(p) + 1) \left(\frac{\log p}{n} \right)^{1-q/2}. \quad (29)$$

Moreover, when $\log p = o(n)$, $s_0(p) \leq Mn^{(1-q)/2}(\log p)^{-(3-q)/2}$ for some constant $M > 0$, the rate in (27) is optimal as we also have the lower bound

$$\inf_{\hat{\mathbf{R}}} \sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}} - \mathbf{R}\|^2 \geq cs_0^2(p) \left(\frac{\log p}{n} \right)^{1-q} \quad (30)$$

$$\inf_{\hat{\mathbf{R}}} \sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}} - \mathbf{R}\|_{\ell_1}^2 \geq cs_0^2(p) \left(\frac{\log p}{n} \right)^{1-q} \quad (31)$$

$$\inf_{\hat{\mathbf{R}}} \sup_{\mathbf{R} \in \mathcal{G}_q^1(s_0(p))} E \|\hat{\mathbf{R}} - \mathbf{R}\|_F^2 \geq cps_0(p) \left(\frac{\log p}{n} \right)^{1-q/2}. \quad (32)$$

Remark 6.1 Cai and Liu (2011) proposed an adaptive thresholding estimator $\hat{\Sigma}^*$ of a sparse covariance matrix Σ . This estimator leads naturally to an estimator $\tilde{\mathbf{R}} = (\tilde{r}_{ij})$ of a sparse correlation matrix \mathbf{R} by normalizing $\hat{\Sigma}^* = (\hat{\sigma}_{ij}^*)$ via $\tilde{r}_{ij} = \hat{\sigma}_{ij}^* (\sigma_{ii}^* \sigma_{jj}^*)^{-1/2}$. The correlation matrix estimator $\tilde{\mathbf{R}}$ has similar properties as the estimator introduced above. For example, $\tilde{\mathbf{R}}$ and $\hat{\mathbf{R}}^*$ achieve the same rate of convergence.

6.3 Optimal Estimation of Sparse Differential Covariance Matrices

Our analysis can also be used for estimation of sparse differential covariance matrices, $\Delta = \Sigma_1 - \Sigma_2$. Define θ_{ijt} as in (3) and its estimate $\hat{\theta}_{ijt}$ as in (8). Similar to the estimation of the

differential correlation matrix $\mathbf{D} = \mathbf{R}_1 - \mathbf{R}_2$, we estimate $\mathbf{\Delta} = \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2$ by adaptive entrywise thresholding. Specifically, we define the thresholding estimator $\hat{\mathbf{\Delta}}^* = (\hat{\delta}_{ij}^*) \in \mathbb{R}^{p \times p}$ by

$$\hat{\delta}_{ij}^* = s_{\gamma_{ij}}(\hat{\sigma}_{ij1} - \hat{\sigma}_{ij2}), \quad 1 \leq i, j \leq p \quad (33)$$

where γ_{ij} is the thresholding level given by

$$\gamma_{ij} = \tau \left(\left(\frac{\log p}{n_1} \hat{\theta}_{ij1} \right)^{1/2} + \left(\frac{\log p}{n_2} \hat{\theta}_{ij2} \right)^{1/2} \right). \quad (34)$$

Same as in the last section, here $s_\lambda(z)$ belongs to the class of thresholding functions satisfying Conditions (C1)-(C3) and the thresholding constant τ can be taken chosen empirically by cross-validation.

We consider the following class of paired covariance matrices with approximately sparse differences, for $0 \leq q < 1$,

$$\mathcal{F}_q(s_0(p)) \triangleq \left\{ (\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) : \mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \succeq 0, \max_{1 \leq i \leq p, t=1,2} \sigma_{iit} \leq B, \max_i \sum_{j=1}^p |\sigma_{ij1} - \sigma_{ij2}|^q \leq s_0(p) \right\}. \quad (35)$$

Under the same conditions as those in Theorems 3.1 and 3.2, a similar analysis can be used to derive the minimax upper and lower bounds. It can be shown that the estimator $\hat{\mathbf{\Delta}}^*$ given in (33) with $\tau > 4$ satisfies

$$\sup_{(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{F}_q(s_0(p))} E \|\hat{\mathbf{\Delta}}^* - (\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\|^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (36)$$

for some constant $C > 0$. Furthermore, the following minimax lower bound holds,

$$\inf_{\hat{\mathbf{\Delta}}} \sup_{(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \in \mathcal{F}_q(s_0(p))} E \|\hat{\mathbf{\Delta}} - (\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\|^2 \geq c s_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (37)$$

for some constant $c > 0$. Equations (36) and (37) together show that the thresholding estimator $\hat{\mathbf{\Delta}}^*$ defined in (33) and (34) is rate-optimal.

6.4 Estimate Differential Cross-Correlation Matrices

In many applications such as phenome-wide association studies (PheWAS) which aims to study the relationship between a set of genomic markers \mathbf{X} and a range of phenotypes \mathbf{Y} , the main focus is on the cross-correlations between the components of \mathbf{X} and those of \mathbf{Y} . That is, the object of interest is a submatrix of the correlation matrix of the random vector $\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. More

specifically, let $\mathbf{X} = (X_1, \dots, X_{p_1})'$ be a p_1 -dimensional random vector and $\mathbf{Y} = (Y_1, \dots, Y_{p_2})'$ be a p_2 -dimensional random vector. In the PheWAS setting, \mathbf{X} may be all phenotypic disease conditions of interest and \mathbf{Y} is a vector of genomic markers.

Suppose we have two independent and identical distributed samples of the (\mathbf{X}, \mathbf{Y}) pairs, one for the case group and one for the control group,

$$\begin{pmatrix} X_k^{(1)} \\ Y_k^{(1)} \end{pmatrix} = \begin{pmatrix} X_{1k}^{(1)} \\ \vdots \\ X_{p_1 k}^{(1)} \\ Y_{1k} \\ \vdots \\ Y_{p_2 k}^{(1)} \end{pmatrix}, \quad k = 1, \dots, n_1; \quad \begin{pmatrix} X_k^{(2)} \\ Y_k^{(2)} \end{pmatrix} = \begin{pmatrix} X_{1k}^{(2)} \\ \vdots \\ X_{p_1 k}^{(2)} \\ Y_{1k} \\ \vdots \\ Y_{p_2 k}^{(2)} \end{pmatrix}, \quad k = 1, \dots, n_2$$

Here for $t = 1, 2$, $(X_k^{(t)T}, Y_k^{(t)T})^T$, $k = 1, \dots, n_t$ are independent and identical distributed samples generated from some distribution with mean μ_t , covariance matrix Σ_t and correlation matrix \mathbf{R}_t given by

$$\mu_t = \begin{pmatrix} \mu_{Xt} \\ \mu_{Yt} \end{pmatrix}, \quad \Sigma_t = \begin{bmatrix} \Sigma_{XXt} & \Sigma_{XYt} \\ \Sigma_{YXt} & \Sigma_{YYt} \end{bmatrix}, \quad \mathbf{R}_t = \begin{bmatrix} \mathbf{R}_{XXt} & \mathbf{R}_{XYt} \\ \mathbf{R}_{YXt} & \mathbf{R}_{YYt} \end{bmatrix}$$

In applications such as PheWAS, it is of special interest to estimate the differential cross-correlation matrix of \mathbf{X} and \mathbf{Y} , i.e. $\mathbf{D}_{XY} = \mathbf{R}_{XY1} - \mathbf{R}_{XY2} \in \mathbb{R}^{p_1 \times p_2}$. Again, we introduce the following set of paired correlation matrices with sparse cross-correlations,

$$\mathcal{G}_q(s_0(p_1, p_2)) = \left\{ (\mathbf{R}_1, \mathbf{R}_2) : \mathbf{R}_1, \mathbf{R}_2 \succeq 0, \text{diag}(\mathbf{R}_1) = \text{diag}(\mathbf{R}_2) = 1; \right. \\ \left. \max_{1 \leq i \leq p_1} \sum_{j=1}^{p_2} |(r_{XY})_{ij1} - (r_{XY})_{ij2}|^q \leq s_0(p_1, p_2) \right\}, \quad 0 \leq q < 1.$$

The thresholding procedure proposed in Section 2 can be applied to estimate \mathbf{D}_{XY} ,

$$(\hat{d}_{XY}^*)_{ij} = s_{\lambda_{ij}}((\hat{r}_{XY})_{ij1} - (\hat{r}_{XY})_{ij2}), \quad 1 \leq i \leq p_1, \quad 1 \leq j \leq p_2 \quad (38)$$

where $\hat{\mathbf{R}}_{XY}$ is sample cross-correlation matrix of X and Y ; λ_{ij} is given by (10). Similar to Theorem 3.1, the following theoretical results hold for the estimator $\hat{\mathbf{D}}_{XY}^* = (\hat{d}_{XY}^*)$.

Proposition 6.2 Suppose $p = p_1 + p_2$, $\log(p) = o(\min(n_1, n_2)^{1/3})$ and (12) and (13) hold. Suppose the thresholding function $s_\lambda(z)$ satisfies Conditions (C1)-(C3). Then $\hat{\mathbf{D}}^*$ defined in (38) with the thresholding constant $\tau > 4$ satisfies

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p_1, p_2))} E \|\hat{\mathbf{D}}_{XY}^* - (\mathbf{R}_{XY1} - \mathbf{R}_{XY2})\|^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (39)$$

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p_1, p_2))} E \|(\hat{\mathbf{D}}_{XY}^* - (\mathbf{R}_{XY1} - \mathbf{R}_{XY2}))^\top\|_{\ell_1}^2 \leq C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \quad (40)$$

$$\sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p_1, p_2))} E \|\hat{\mathbf{D}}_{XY}^* - (\mathbf{R}_{XY1} - \mathbf{R}_{XY2})\|_F^2 \leq Cp(s_0(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q/2} \quad (41)$$

for some constant $C > 0$ that does not depend on n_1, n_2 or p .

The proof of Proposition 6.2 is similar to that of Theorem 3.1 by analyzing the block $\hat{\mathbf{D}}_{XY} - (\mathbf{R}_{XY1} - \mathbf{R}_{XY2})$ instead of the whole matrix $\mathbf{D}^* - (\mathbf{R}_1 - \mathbf{R}_2)$. We omit the detailed proof here.

References

- Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W.-K., Aebersold, R., Keogh, M.-C., Krogan, N. J., and Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Sci. Signal.*, 330(6009):1385–1389.
- Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, 12(1):56–68.
- Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.*, 5(2):101–113.
- Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 26:879–921.
- Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of American Statistical Association*, 106:672–684.

- Cai, T. T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of American Statistical Association*, 108:265–277.
- Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40:2014–2042.
- Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38:2118–2144.
- Cai, T. T. and Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40:2389–2420.
- Carter, S. L., Brechbühler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250.
- de la Fuente, A. (2010). From “differential expression” to “differential networking” identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26:326–333.
- Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G. A., Ádány, R., Aromaa, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302.
- Fukushima, A. (2013). Diffcorr: an r package to analyze and visualize differential correlations in biological networks. *Gene*, 518(1):209–214.
- Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusk, A. J., and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome*, 18(6-7):463–472.
- Ghellal, A., Li, C., Hayes, M., Byrne, G., Bundred, N., and Kumar, S. (2000). Prognostic significance of tgf beta 1 and tgf beta 3 in human breast carcinoma. *Anticancer Rec.*, 20:4413–4418.

- Hudson, N. J., Reverter, A., and Dalrymple, B. P. (2009). A differential wiring analysis of expression data correctly identifies the gene containing the causal mutation. *PLoS Comput. Biol.*, 5(5):e1000382.
- Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Molecular systems biology*, 8(1):565.
- Klopocki, E., Kristiansen, G., Wild, P. J., Klamann, I., Castanos-Velez, E., Singer, G., Sthir, R., Simon, R., Sauter, G., Leibiger, H., Essers, L., Weber, B., Hermann, K., Rosenthal, A., Hartmann, A., and Dahl, E. (2004). Loss of sfrp1 is associated with breast cancer progression and poor prognosis in early stage tumors. *Int. J. Oncol.*, 25:641–649.
- Kostka, D. and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Bioinformatics*, 20(suppl 1):i194–i199.
- Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics*, 20:3146–3155.
- Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome research*, 14(6):1085–1094.
- Li, J., Chen, S. X., et al. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.
- Margheri, F., Schiavone, N., Papucci, L., Magnelli, L., Serrat, S., Chill, A., Laurenzana, A., Bianchini, F., Calorini, L., Torre, E., Dotor, J., Feijoo, E., Fibbi, G., and Del Rosso, M. (2012). Gdf5 regulates tgf-dependent angiogenesis in breast carcinoma mcf-7 cells: In vitro and in vivo control by anti-tgf peptides. *PloS ONE*, 7:e50342.
- Rothman, A., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104:177–186.
- Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis*, 51(12):6535–6542.

- Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., and Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, 440(7084):676–679.
- Shedden, K. and Taylor, J. (2005). Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In *Methods of Microarray Data Analysis*, pages 121–131. Springer.
- Shy, B., Wu, C., Khramtsova, G., Zhang, J., Olopade, O., Goss, K., and Merrill, B. (2013). Regulation of tcf7l1 dna binding and protein stability as principal mechanisms of wnt/b-catenin signaling. *Cell Reports*, 4:1–9.
- Srivastava, M. S. and Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101(6):1319–1329.
- van de Vijver, M., He, Y., van’t Veer, L., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., vander Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347:1999–2009.
- Xu, Y., Liu, X., Guo, F., Ning, Y., Zhi, X., Wang, X., Chen, S., Yin, L., and Li, X. (2012). Effect of estrogen sulfation by sult1e1 and papss on the development of estrogen-dependent cancers. *Cancer science*, 103:1000–1009.
- Zafrakas, M., Chorovicer, M., Klamann, I., Kristiansen, G., Wild, P.-J., Heindrichs, U., Knüchel, R., and Dahl, E. (2006). Systematic characterisation of gabrp expression in sporadic breast cancer and normal breast tissue. *International Journal of Cancer*, 118(6):1453–1459.
- Zhang, J., Li, J., and Deng, H. (2008). Class-specific correlations of gene expressions: identification and their effects on clustering analyses. *The American Journal of Human Genetics*, 83(2):269–277.

Appendix: Proofs

We prove the main theorems in the Appendix. Throughout the Appendix, we denote by C a constant which does not depend on p, n_1 and n_2 , and may vary from place to place.

Proof of Theorem 3.1 To prove this theorem, we consider the following three events separately,

$$A_1 = \left\{ \max_{ijt} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\hat{\theta}_{ijt} \log p / n_t)^{1/2}} \leq \frac{\tau}{4} + 3, \quad \text{and} \quad \max_{ijt} \frac{|\hat{\theta}_{ijt} - \theta_{ijt}|}{\sigma_{iit} \sigma_{jjt}} \leq \varepsilon \right\} \quad (42)$$

$$A_2 = \left\{ \max_{ijt} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\hat{\theta}_{ijt} \log p / n_t)^{1/2}} > \frac{\tau}{4} + 3, \quad \max_{ijt} \frac{|\hat{\theta}_{ijt} - \theta_{ijt}|}{\sigma_{iit} \sigma_{jjt}} \leq \varepsilon, \right. \\ \left. \text{and} \quad \max_{ijt} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\sigma_{iit} \sigma_{jjt})^{1/2}} \leq \min(0.5, C_1 C_3) \right\} \quad (43)$$

$$A_3 = (A_1 \cup A_2)^c. \quad (44)$$

Here ε is the fixed constant which satisfies $0 < \varepsilon < \nu_0/2$ where ν_0 is introduced in (13); C_1 and C_3 are constants which do not depend on p, n_1, n_2 and shall be specified later in Lemma 6.1.

1. First we would like to show that under the event A_1 ,

$$\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \leq C s_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}, \quad (45)$$

$$\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 \leq C s_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}, \quad (46)$$

$$\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 \leq C p s_0(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q/2}. \quad (47)$$

In fact,

$$E Y_i^{(t)4} \leq \frac{2E \exp(\eta Y_i^{(t)2})}{\eta^2} \leq \frac{2K}{\eta^2}$$

for all $1 \leq i \leq p$, so

$$\begin{aligned} \theta_{ijt} &= \text{Var}(X_i^{(t)} - \mu_i)(X_j^{(t)} - \mu_j) \leq E(X_i^{(t)} - \mu_i)^2 (X_j^{(t)} - \mu_j)^2 \\ &\leq \left(E(X_i^{(t)} - \mu_i)^4 E(X_j^{(t)} - \mu_j)^4 \right)^{1/2} = \sigma_{iit} \sigma_{jjt} \left(E Y_i^{(t)4} E Y_j^{(t)4} \right)^{1/2} \leq C \sigma_{iit} \sigma_{jjt}, \\ \theta_{ijt} &= \text{Var}(Y_i Y_j) \cdot \sigma_{iit} \sigma_{jjt} \stackrel{(13)}{\geq} \nu_0 \sigma_{iit} \sigma_{jjt}. \end{aligned} \quad (48)$$

So by the definition of A_1 , we have

$$\hat{\theta}_{ijt} \leq \theta_{ijt} + |\hat{\theta}_{ijt} - \theta_{ijt}| \leq C\sigma_{iit}\sigma_{jzt}, \text{ for all } i, j, t, \quad (49)$$

$$\hat{\theta}_{ijt} \geq \theta_{ijt} - |\hat{\theta}_{ijt} - \theta_{ijt}| \geq \nu_0\sigma_{iit}\sigma_{jzt} - \varepsilon\sigma_{iit}\sigma_{jzt} \geq \frac{\nu_0}{2}\sigma_{iit}\sigma_{jzt}. \quad (50)$$

Hence,

$$\left| \frac{\hat{\sigma}_{iit}}{\sigma_{iit}} - 1 \right| \leq \frac{|\hat{\sigma}_{iit} - \sigma_{iit}|}{\sigma_{iit}} \stackrel{(42)}{\leq} \frac{\tau/4 + 3}{\sigma_{iit}} \left(\log p \frac{\hat{\theta}_{iit}}{n_t} \right)^{1/2} \stackrel{(49)}{\leq} C \left(\frac{\log p}{n_t} \right)^{1/2} \quad (51)$$

$$\left| \frac{\sigma_{iit}}{\hat{\sigma}_{iit}} - 1 \right| \leq \frac{|\hat{\sigma}_{iit} - \sigma_{iit}|}{\hat{\sigma}_{iit}} \stackrel{(42)}{\leq} (\tau/4 + 3) \left(\frac{\log p}{n_t} \right)^{1/2} \frac{\hat{\theta}_{iit}^{1/2}}{\hat{\sigma}_{iit}} \quad (52)$$

Suppose $x = \sigma_{iit}/\hat{\sigma}_{iit}$, $y = \sigma_{jzt}/\hat{\sigma}_{jzt}$. By (51) and $\left(\frac{\log p}{n_t} \right)^{1/2} \rightarrow 0$, we have $\max\{|x - 1|, |y - 1|\} \leq C \left(\frac{\log p}{n_t} \right)^{1/2}$ when n_t is large enough. Thus for large n_t , we obtain

$$\begin{aligned} \left| \left(\frac{\sigma_{iit}\sigma_{jzt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jzt}} \right)^{1/2} - 1 \right| &= |(xy)^{1/2} - 1| = \frac{|xy - 1|}{(xy)^{1/2} + 1} \leq \frac{|x - 1| + |y - 1|}{2 - \max(|x - 1|, |y - 1|)} \\ &\leq \frac{\max(1, x)}{2 - \max(|x - 1|, |y - 1|)} (|x - 1| + |y - 1|) \\ &\leq \left(\frac{1}{2} + O \left(\left(\frac{\log p}{n_t} \right)^{1/2} \right) \right) (|x - 1| + |y - 1|). \end{aligned} \quad (53)$$

It then follows from the assumption $\log p = o(n_t^{1/3})$ that for large n_t ,

$$\hat{\xi}_{ijt} = \frac{\hat{\theta}_{ijt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jzt}} \stackrel{(49)}{\leq} \frac{C\sigma_{iit}\sigma_{jzt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jzt}} \stackrel{(51)}{\leq} C \quad (54)$$

and

$$\begin{aligned}
|\hat{r}_{ijt} - r_{ijt}| &= \left| \frac{\hat{\sigma}_{ijt}}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} - \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} \right| \leq \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\hat{\sigma}_{iit}\hat{\sigma}_{jjt})^{1/2}} + \frac{|\sigma_{ijt}|}{(\sigma_{iit}\sigma_{jjt})^{1/2}} \left| \left(\frac{\sigma_{iit}\sigma_{jjt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jjt}} \right)^{1/2} - 1 \right| \\
&\stackrel{(42)(53)}{\leq} \left(\frac{\tau}{4} + 3 \right) \left(\frac{\log p}{n_t} \frac{\hat{\theta}_{ijt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jjt}} \right)^{1/2} + |r_{ijt}| \left(\frac{1}{2} + O \left(\left(\frac{\log p}{n_t} \right)^{1/2} \right) \right) \left(\left| \frac{\sigma_{iit}}{\hat{\sigma}_{iit}} - 1 \right| + \left| \frac{\sigma_{jjt}}{\hat{\sigma}_{jjt}} - 1 \right| \right) \\
&\stackrel{(52)}{\leq} \left(\frac{\tau}{4} + 3 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\left(\frac{\hat{\theta}_{ijt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jjt}} \right)^{1/2} + |r_{ijt}| \left(\frac{1}{2} + O \left(\left(\frac{\log p}{n_t} \right)^{1/2} \right) \right) \left(\frac{\hat{\theta}_{iit}^{1/2}}{\hat{\sigma}_{iit}} + \frac{\hat{\theta}_{jjt}^{1/2}}{\hat{\sigma}_{jjt}} \right) \right) \\
&\leq \left(\frac{\tau}{2} + 2 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|r_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \right) \\
&\leq \left(\frac{\tau}{2} + 2 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|\hat{r}_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \right) \\
&\quad + |\hat{r}_{ijt} - r_{ijt}| \left(\frac{\tau}{4} + 1 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \\
&\stackrel{(54)}{\leq} \left(\frac{\tau}{2} + 2 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|\hat{r}_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \right) + C \left(\frac{\log p}{n_t} \right)^{1/2} |\hat{r}_{ijt} - r_{ijt}|.
\end{aligned}$$

We shall note the difference between $\frac{|r_{ijt}|}{2}$ and $\frac{|\hat{r}_{ijt}|}{2}$ above. Next, we rearrange the inequality above and write it into an inequality for $|\hat{r}_{ijt} - r_{ijt}|$,

$$\begin{aligned}
|\hat{r}_{ijt} - r_{ijt}| &\leq \frac{\left(\frac{\tau}{2} + 2 \right) \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|\hat{r}_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \right)}{1 - C \left(\frac{\log p}{n_t} \right)^{1/2}} \\
&\leq \tau \left(\frac{\log p}{n_t} \right)^{1/2} \left(\hat{\xi}_{ijt}^{1/2} + \frac{|\hat{r}_{ijt}|}{2} \left(\hat{\xi}_{iit}^{1/2} + \hat{\xi}_{jjt}^{1/2} \right) \right) \stackrel{(10)}{=} \lambda_{ijt}.
\end{aligned} \tag{55}$$

(55) implies

$$|(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \leq \lambda_{ij1} + \lambda_{ij2} = \lambda_{ij} \text{ holds for all } 1 \leq i, j \leq p \tag{56}$$

Next, by (56) and (C1) and (C3) of $s_\lambda(z)$,

$$\begin{aligned}
|s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| &\leq |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2})| + |(r_{ij1} - r_{ij2})| \\
&\leq (1 + c)|r_{ij1} - r_{ij2}|,
\end{aligned} \tag{57}$$

$$\begin{aligned}
&|s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \\
&\leq |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (\hat{r}_{ij1} - \hat{r}_{ij2})| + |(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \leq 2\lambda_{ij},
\end{aligned} \tag{58}$$

which implies

$$|s_{r_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \leq (2\lambda_{ij})^{1-q}(1+c)^q|r_{ij1} - r_{ij2}|^q, \tag{59}$$

$$|s_{r_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})|^2 \leq (2\lambda_{ij})^{2-q}(1+c)^q|r_{ij1} - r_{ij2}|^q, \quad (60)$$

where $0 \leq q < 1$. Hence,

$$\begin{aligned} & \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1} \\ &= \max_i \sum_{j=1}^p |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \\ &\stackrel{(59)}{\leq} \max_i 2^{1-q}(1+c)^q \sum_{j=1}^p \lambda_{ij}^{1-q} |r_{ij1} - r_{ij2}|^q \\ &\stackrel{(10)}{\leq} \max_i 2^{1-q}(1+c)^q \sum_{j=1}^p \left\{ \tau^{1-q} (\log p)^{(1-q)/2} \right. \\ &\quad \times \left(\frac{\hat{\xi}_{ij1}^{1/2} + |\hat{r}_{ij1}|(\hat{\xi}_{ii1}^{1/2} + \hat{\xi}_{jj1}^{1/2})/2}{n_1^{1/2}} + \frac{\hat{\xi}_{ij2}^{1/2} + |\hat{r}_{ij2}|(\hat{\xi}_{ii2}^{1/2} + \hat{\xi}_{jj2}^{1/2})/2}{n_2^{1/2}} \right)^{1-q} |r_{ij1} - r_{ij2}|^q \Big\} \\ &\stackrel{(54)}{\leq} \max_i C \sum_{j=1}^p \left\{ (\log p)^{(1-q)/2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{(1-q)/2} |r_{ij1} - r_{ij2}|^q \right\} \\ &\leq C s_0(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{(1-q)/2}. \end{aligned}$$

which yields to (46). (45) also holds due to the fact that $\|A\|_2 \leq \|A\|_{L_1}$ for any symmetric matrix A . Similarly,

$$\begin{aligned} & \left\| \hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2) \right\|_F^2 \\ &= \sum_{i=1}^p \sum_{j=1}^p |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})|^2 \\ &\stackrel{(60)}{\leq} 2^{2-q}(1+c)^q \sum_{i=1}^p \sum_{j=1}^p \lambda_{ij}^{2-q} |r_{ij1} - r_{ij2}|^q \\ &\stackrel{(10)}{\leq} 2^{2-q}(1+c)^q \sum_{i=1}^p \sum_{j=1}^p \left\{ \tau^{2-q} (\log p)^{(2-q)/2} \right. \\ &\quad \times \left(\frac{\hat{\xi}_{ij1}^{1/2} + |\hat{r}_{ij1}|(\hat{\xi}_{ii1}^{1/2} + \hat{\xi}_{jj1}^{1/2})/2}{n_1^{1/2}} + \frac{\hat{\xi}_{ij2}^{1/2} + |\hat{r}_{ij2}|(\hat{\xi}_{ii2}^{1/2} + \hat{\xi}_{jj2}^{1/2})/2}{n_2^{1/2}} \right)^{2-q} |r_{ij1} - r_{ij2}|^q \Big\} \\ &\stackrel{(54)}{\leq} Cp \max_i \sum_{j=1}^p \left\{ (\log p)^{(2-q)/2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{(2-q)/2} |r_{ij1} - r_{ij2}|^q \right\} \\ &\leq Cp s_0(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q/2}. \end{aligned}$$

which implies (47).

2. For A_2 , we wish to prove,

$$\int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 dP \leq C(p^{-\tau/4+1} \log p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (61)$$

$$\int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 dP \leq C(p^{-\tau/4+1} \log p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (62)$$

$$\int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 dP \leq C(p^{-\tau/4+1} \log p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (63)$$

In order to prove these probability bounds, we introduce the following lemma, which revealed the relationship between $\hat{\theta}_{ijt}, \theta_{ijt}$ and $\hat{\sigma}_{ijt}, \sigma_{ijt}$.

Lemma 6.1 *For any $\tau > 0$,*

$$\text{pr} \left(\max_{i,j,t} |\hat{\sigma}_{ijt} - \sigma_{ijt}| > (\tau/4 + 3) \left(\hat{\theta}_{ijt} \log p / n_t \right)^{1/2} \right) \leq C(\log p)^{-1/2} p^{-\tau/4-1}, \quad (64)$$

There exist constants C_1, C_2, C_3 which do not depend on p, n_1, n_2 such that

$$\text{pr} \left(\max_{i,j} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\sigma_{iit}\sigma_{jtt})^{1/2}} > C_1 x \right) \leq C_2 p^2 (\exp(-n_t x^2)), \quad \text{for all } 0 < x \leq C_3, t = 1, 2; \quad (65)$$

For any $\varepsilon > 0$ and $M > 0$,

$$\text{pr} \left(\max_{i,j,t} |\hat{\theta}_{ijt} - \theta_{ijt}| / (\sigma_{iit}\sigma_{jtt}) > \varepsilon \right) \leq C p^{-M} (1/n_1 + 1/n_2) \quad (66)$$

The proof of Lemma 6.1 is given later. Note that (64) immediately leads to

$$\text{pr}(A_2) \leq C(\log p)^{-1/2} p^{-\tau/4-1}. \quad (67)$$

By the definition of A_2 (43), we still have (49). Besides, by the definition of A_2 , $\frac{|\hat{\sigma}_{iit} - \sigma_{iit}|}{\sigma_{iit}} \leq 0.5$, which leads to $\hat{\sigma}_{iit} \geq 0.5\sigma_{iit}$. Thus,

$$\hat{\xi}_{ijt} = \frac{\hat{\theta}_{ijt}}{\hat{\sigma}_{iit}\hat{\sigma}_{jtt}} \leq \frac{C\sigma_{iit}\sigma_{jtt}}{(0.5\sigma_{iit})(0.5\sigma_{jtt})} = 4C. \quad (68)$$

For convenience, we denote the random variable

$$T = \max_{ijt} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\sigma_{iit}\sigma_{jtt})^{1/2}}. \quad (69)$$

Under A_2 , we have $T \leq 0.5$. Then for all $1 \leq i, j \leq p, t = 1, 2$,

$$\begin{aligned}
\hat{r}_{ijt} - r_{ijt} &= \frac{\hat{\sigma}_{ijt}}{(\hat{\sigma}_{iit}\hat{\sigma}_{jtt})^{1/2}} - \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jtt})^{1/2}} \\
&= \frac{\frac{\hat{\sigma}_{ijt}}{(\sigma_{iit}\sigma_{jtt})^{1/2}}}{(\hat{\sigma}_{iit}/\sigma_{iit})^{1/2} \times (\hat{\sigma}_{jtt}/\sigma_{jtt})^{1/2}} - \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jtt})^{1/2}} \\
&\leq \frac{\frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jtt})^{1/2}} + T}{(\sigma_{iit}/\sigma_{iit} - T)^{1/2} \times (\sigma_{jtt}/\sigma_{jtt} - T)^{1/2}} - \frac{\sigma_{ijt}}{(\sigma_{iit}\sigma_{jtt})^{1/2}} \\
&= \frac{r_{ijt} + T}{1 - T} - r_{ijt} \\
&\leq (1 + 2T)(r_{ijt} + T) - r_{ijt} \\
&\leq 4T.
\end{aligned}$$

Similarly calculation also leads to $\hat{r}_{ijt} - r_{ijt} \geq -4T$. Then, by (C3) of $s_{\lambda_{ij}}(z)$,

$$\begin{aligned}
\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 &= \max_i \left(\sum_{j=1}^p |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \right)^2 \\
&\leq \max_i \left(\sum_{j=1}^p (|s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (\hat{r}_{ij1} - \hat{r}_{ij2})| + |(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})|) \right)^2 \quad (70) \\
&\leq \max_i \left(\sum_{j=1}^p (\lambda_{ij} + 8T) \right)^2 \stackrel{(10)(68)}{\leq} Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} + T^2 \right).
\end{aligned}$$

In addition, due to $\|\cdot\|_{\ell_1} \geq \|\cdot\|$, we also have $\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \leq Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} + T^2 \right)$.

Similarly,

$$\begin{aligned}
&\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 \\
&= \sum_{i=1}^p \sum_{j=1}^p |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})|^2 \\
&\leq 2 \sum_{i=1}^p \sum_{j=1}^p (|s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (\hat{r}_{ij1} - \hat{r}_{ij2})|^2 + |(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})|^2) \quad (71) \\
&\leq Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} + T^2 \right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 dP \\
& \stackrel{(70)}{\leq} \int_{A_2} Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} + T^2 \right) dP \\
& \leq Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) \text{pr}(A_2) + Cp^2 \int_0^{\min(0.5, C_1 C_3)} 2x \text{pr}(\{T \geq x\} \cap A_2) dx \\
& \leq Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) \text{pr}(A_2) + Cp^2 \int_0^{C_1(M \log p(1/n_1 + 1/n_2))^{1/2}} 2x \text{pr}(A_2) dx \\
& \quad + \int_{C_1(M \log p(1/n_1 + 1/n_2))^{1/2}}^{\min(0.5, C_1 C_3)} 2x \text{pr}(T \geq x) dx \\
& \stackrel{(65)}{\leq} Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) \text{pr}(A_2) + Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) \text{pr}(A_2) \\
& \quad + \int_{C_1(M \log p(1/n_1 + 1/n_2))^{1/2}}^{+\infty} 2xC_2 \left(\exp(-n_1(x/C_1)^2) + \exp(-n_2(x/C_1)^2) \right) dx \\
& \leq Cp^2 \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) \text{pr}(A_2) \\
& \quad + C \left(\frac{1}{n_1} \exp(-n_1(x/C_1)^2) + \frac{1}{n_2} \exp(-n_2(x/C_1)^2) \right) \Big|_{+\infty}^{C_1(M \log p(1/n_1 + 1/n_2))^{1/2}} \\
& \stackrel{(67)}{\leq} Cp^{-\tau/4+1} \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) + Cp^{-M} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)
\end{aligned} \tag{72}$$

Similarly, we have

$$\begin{aligned}
& \int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 dP \leq Cp^{-\tau/4+1} \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) + Cp^{-M} \left(\frac{1}{n_1} + \frac{1}{n_2} \right), \\
& \int_{A_2} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 dP \leq Cp^{-\tau/4+1} \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right) + Cp^{-M} \left(\frac{1}{n_1} + \frac{1}{n_2} \right),
\end{aligned}$$

which finishes the proof of (61), (62) and (63) when we choose $M > \tau/4 - 1$.

3. For A_3 , (66) and $\log p = o(n^{1/3})$ leads to

$$\begin{aligned}
\text{pr}(A_3) & \leq \text{pr} \left(\max_{ijt} \frac{|\hat{\theta}_{ijt} - \theta_{ijt}|}{\sigma_{iit}\sigma_{jjt}} > \varepsilon \right) + \text{pr} \left(\max_{ijt} \frac{|\hat{\sigma}_{ijt} - \sigma_{ijt}|}{(\sigma_{iit}\sigma_{jjt})^{1/2}} > \min(0.5, C_1 C_3) \right) \\
& \leq Cp^{-M}(1/n_1 + 1/n_2) + C_2 p^2 \left(\exp(-n_1 \min(\frac{1}{2C_1}, C_3)^2) + \exp(-n_2 \min(\frac{1}{2C_1}, C_3)^2) \right) \\
& = Cp^{-M}(1/n_1 + 1/n_2)
\end{aligned} \tag{73}$$

Besides, since r_{ijt}, \hat{r}_{ijt} are the population and sample correlations, $|r_{ijt}| \leq 1, |\hat{r}_{ijt}| \leq 1$. By (C1) of thresholding $s_\lambda(z)$, we have $|s_\lambda(x) - x| \leq c|x|$ for all $x \in \mathbb{R}$. Thus,

$$\begin{aligned} |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| &\leq |r_{ij1}| + |r_{ij2}| + |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2})| \\ &\leq 2 + c|\hat{r}_{ij1} - \hat{r}_{ij2}| \leq 2 + 2c \end{aligned}$$

which yields

$$\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 = \max_i \left(\sum_{j=1}^p |s_{\lambda_{ij}}(\hat{r}_{ij1} - \hat{r}_{ij2}) - (r_{ij1} - r_{ij2})| \right)^2 \leq (2 + 2c)^2 p^2 \quad (74)$$

Similarly, $\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \leq (2 + 2c)^2 p^2$, $\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 \leq (2 + 2c)^2 p^2$. Therefore,

$$\int_{A_3} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 dP \stackrel{(73)}{\leq} Cp^{-M+2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (75)$$

$$\int_{A_3} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_{\ell_1}^2 dP \stackrel{(73)}{\leq} Cp^{-M+2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (76)$$

$$\int_{A_3} \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|_F^2 dP \stackrel{(73)}{\leq} Cp^{-M+2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \quad (77)$$

Finally, we combine the situations of A_1, A_2 and A_3 . When $\tau > 4$ and $M > 2$, we have

$$\begin{aligned} E\|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 &= \left(\int_{A_1} + \int_{A_2} + \int_{A_3} \right) \|\hat{\mathbf{D}}^* - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 dP \\ &\stackrel{(45)(61)(75)}{\leq} C(s_0^2(p) + 1) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q} \end{aligned} \quad (78)$$

which has proved (14). (15) and (16) can be proved similarly by (46), (62), (76) and (47), (63), (77). \square

Proof of Lemma 6.1. (64) is directly from (25) in Cai and Liu (2011). For (66), the proof is essentially the same as the proof of (26) in Cai and Liu (2011) as long as we use $x = ((M + 2) \log p + \log n)^{1/2}$ in stead of $x = ((M + 2) \log p)^{1/2}$ in their proof. Now we mainly focus on the proof of (65). Without loss of generality, we can translate X and assume that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$. Note that we have the following formulation,

$$\frac{\hat{\sigma}_{ijt} - \sigma_{ijt}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} = \frac{1}{n_t} \sum_{k=1}^{n_t} \frac{(X_{ik}^{(t)} X_{jk}^{(t)} - \sigma_{ijt})}{(\sigma_{iit}\sigma_{jjt})^{1/2}} - \frac{\bar{X}_i^{(t)} \bar{X}_j^{(t)}}{(\sigma_{iit}\sigma_{jjt})^{1/2}} = \left(\frac{1}{n_t} \sum_{k=1}^{n_t} (Y_{ik}^{(t)} Y_{jk}^{(t)} - r_{ijt}) - \bar{Y}_i^{(t)} \bar{Y}_j^{(t)} \right) \quad (79)$$

Since

$$\begin{aligned}
& E(Y_i^{(t)}Y_j^{(t)} - r_{ijt})^2 e^{\frac{\eta}{2}|Y_i^{(t)}Y_j^{(t)} - r_{ijt}|} \\
& \leq \frac{4}{\eta^2} E e^{\eta|Y_i^{(t)}Y_j^{(t)} - r_{ijt}|} \leq \frac{4}{\eta^2} E e^{\eta(Y_i^{(t)}Y_j^{(t)} - r_{ijt})} + \frac{4}{\eta^2} E e^{-\eta(Y_i^{(t)}Y_j^{(t)} - r_{ijt})} \\
& \leq \frac{8}{\eta^2} \left(E e^{\eta|Y_i^{(t)}|^2} + E e^{\eta|Y_j^{(t)}|^2} \right) e^{|\eta r_{ijt}|} \leq C_4
\end{aligned}$$

where C_4 is a constant which does not depend on n_1, n_2, p . Thus, we set $\bar{B}_n^2 = n_t C_1$; based on lemma 1 in Cai and Liu (2011), we have

$$\text{pr} \left(\left| \frac{1}{n_t} \sum_{k=1}^{n_t} (Y_{ik}^{(t)} Y_{jk}^{(t)} - r_{ijt}) \right| \geq C_{\eta/2} C_4^{1/2} x \right) \leq \exp(-n_t x^2). \quad (80)$$

for all $0 < x \leq C_1^{1/2}$, where $C_{\eta/2} = \eta/2 + 2/\eta$. Next for $\bar{Y}_i^{(t)}$, we similarly apply Lemma 1 in Cai and Liu (2011) and get

$$\text{pr} \left(|\bar{Y}_i^{(t)}| \geq C_5 x \right) \leq \exp(-n_t x^2) \quad (81)$$

for all $0 < x \leq C_5^{1/2}$. Combining (80) and (81),

$$\text{pr} \left(\max_{ij} \left| \frac{1}{n_t} \sum_{k=1}^{n_t} (Y_{ik}^{(t)} Y_{jk}^{(t)} - r_{ijt}) \right| \leq C_{\eta/2} C_4^{1/2} x \quad \text{and} \quad \max_{i,t} |\bar{Y}_i^{(t)}| \leq C_5 x \right) \leq 1 - 2p^2 \exp(-n_t x^2) \quad (82)$$

for all $0 < x \leq \min(C_1^{1/2}, C_5^{1/2})$. Finally, (79) and (82) yield (65). \square

Proof of Theorem 3.2. Without loss of generality, we assume $n_1 \leq n_2$. For $(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))$, set $\Sigma_2 = \mathbf{R}_2 = I_{p \times p}$ and we have already known this information. The estimation of sparse difference immediately becomes the estimation of the sparse correlation matrix \mathbf{R}_1 . Then the lower bound result for estimating single sparse covariance matrix can be used to prove this theorem.

We follow the idea of Cai and Zhou (2012) and define the set of diagonal-1 covariance matrices as

$$\mathcal{F}_q(s_0(p)) = \left\{ \Sigma : \Sigma \succeq 0, \text{diag}(\Sigma) = 1, \max_i \sum_{j=1}^p |\sigma_{ij}|^q \leq s_0(p) \right\}.$$

We have $\{(\mathbf{R}_1, \mathbf{I}) : \mathbf{R}_1 \in \mathcal{F}_q(s_0(p))\} \subseteq \mathcal{G}_q(s_0(p))$. Besides, the proof of Theorem 2 in Cai and Zhou (2012) shows that

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{F}_q(s_0(p))} E \|\hat{\Sigma} - \Sigma\|^2 \geq C s_0(p) \left(\frac{\log p}{n} \right)^{1-q} \quad (83)$$

Since the correlation matrix equals to covariance matrix (i.e. $\mathbf{R} = \mathbf{\Sigma}$) when $\text{diag}(\mathbf{\Sigma}) = 1$, then

$$\begin{aligned}
& \inf_{\hat{\mathbf{D}}} \sup_{(\mathbf{R}_1, \mathbf{R}_2) \in \mathcal{G}_q(s_0(p))} E \|\hat{\mathbf{D}} - (\mathbf{R}_1 - \mathbf{R}_2)\|^2 \\
& \geq \inf_{\hat{\mathbf{D}}} \sup_{(\mathbf{R}_1, \mathbf{I}): \mathbf{R}_1 \in \mathcal{F}_q(s_0(p))} E \|\hat{\mathbf{D}} - (\mathbf{R}_1 - \mathbf{I})\|^2 \\
& \geq \inf_{\hat{\mathbf{R}}_1} \sup_{\mathbf{R}_1 \in \mathcal{F}_q(s_0(p))} E \|\hat{\mathbf{R}}_1 - \mathbf{R}_1\|^2 \\
& \geq \inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathbf{\Sigma}_1 \in \mathcal{F}_q(s_0(p)), \text{diag}(\mathbf{\Sigma}_1)=1} E \|\hat{\mathbf{\Sigma}}_1 - \mathbf{\Sigma}_1\| \\
& \geq C s_0^2(p) \left(\frac{\log p}{n_1} \right)^{1-q} \geq \frac{C}{2} s_0^2(p) \left(\frac{\log p}{n_1} + \frac{\log p}{n_2} \right)^{1-q}
\end{aligned} \tag{84}$$

which implies (17). By $\|\cdot\|_{\ell_1} \geq \|\cdot\|$ for symmetric matrices, (18) also follow immediately.

Similarly, (19) follows from Theorem 4 of Cai and Zhou (2012). \square

Proof of Proposition 6.1. The proof of Proposition 6.1 is similar to Theorem 3.1. For the upper bound, again, we split the whole events into three,

$$A_1 = \left\{ \max_{ij} \frac{|\hat{\sigma}_{ij} - \sigma_{ij}|}{\left(\log p \hat{\theta}_{ij} / n \right)^{1/2}} \leq \tau/4 + 3, \quad \text{and} \quad \max_{ij} \frac{|\hat{\theta}_{ij} - \theta_{ij}|}{\sigma_{ii} \sigma_{jj}} \leq \varepsilon \right\}, \tag{85}$$

$$A_2 = \left\{ \max_{ij} \frac{|\hat{\sigma}_{ij} - \sigma_{ij}|}{\left(\log p \hat{\theta}_{ij} / n \right)^{1/2}} > \tau/4 + 3, \quad \max_{ij} \frac{|\hat{\theta}_{ij} - \theta_{ij}|}{\sigma_{ii} \sigma_{jj}} \leq \varepsilon \right. \tag{86}$$

$$\begin{aligned}
& \left. \text{and} \quad \max_{ij} \frac{|\hat{\sigma}_{ij} - \sigma_{ij}|}{(\sigma_{ii} \sigma_{jj})^{1/2}} \leq \min(0.5, C_1 C_3) \right\} \\
& A_3 = (A_1 \cup A_2)^c.
\end{aligned} \tag{87}$$

Here ε is the fixed constant which satisfies $0 < \varepsilon < \nu_0/2$ where ν_0 was introduced in (13); C_1, C_3 are constants specified in Lemma 6.1. Similarly to the proof of Theorem 3.1, we can prove the following statements.

1. Under A_1 ,

$$\begin{aligned}
\|\hat{\mathbf{R}}^* - \mathbf{R}\|^2 & \leq C s_0^2(p) \left(\frac{\log p}{n} \right)^{1-q}, \\
\|\hat{\mathbf{R}}^* - \mathbf{R}\|_{\ell_1}^2 & \leq C s_0^2(p) \left(\frac{\log p}{n} \right)^{1-q}, \\
\|\hat{\mathbf{R}}^* - \mathbf{R}\|_F^2 & \leq C s_0(p) \left(\frac{\log p}{n} \right)^{1-q/2}.
\end{aligned}$$

2. For A_2 ,

$$\begin{aligned}\int_{A_2} \|\hat{\mathbf{R}}^* - \mathbf{R}\|^2 dP &\leq C(p^{-\tau/4+1} \log p) \frac{1}{n} \\ \int_{A_2} \|\hat{\mathbf{R}}^* - \mathbf{R}\|_{\ell_1}^2 dP &\leq C(p^{-\tau/4+1} \log p) \frac{1}{n} \\ \int_{A_2} \|\hat{\mathbf{R}}^* - \mathbf{R}\|_F^2 dP &\leq C(p^{-\tau/4+1} \log p) \frac{1}{n}\end{aligned}$$

3. For A_3 ,

$$\begin{aligned}\int_{A_3} \|\hat{\mathbf{R}}^* - \mathbf{R}\|^2 dP &\leq C \frac{p^{-M+2}}{n} \\ \int_{A_3} \|\hat{\mathbf{R}}^* - \mathbf{R}\|_{\ell_1}^2 dP &\leq C \frac{p^{-M+2}}{n} \\ \int_{A_3} \|\hat{\mathbf{R}}^* - \mathbf{R}\|_F^2 dP &\leq C \frac{p^{-M+2}}{n}\end{aligned}$$

The rest of proof, including the lower bound results, are omitted here as they are essentially the same as Theorem 3.1. \square